NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

TOWARDS REENGINEERING THE UNITED STATES NAVY ENLISTED MANPOWER AND PERSONNEL SYSTEMS – A DATA WAREHOUSE APPROACH

by

Douglas J. Conde
Cassandra A. Crownover

September 2000

Thesis Co-Advisors: Daniel R. Dolk
Julie Filizetti

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2000	Master's Thesis

4. TITLE AND SUBTITLE
Towards Reengineering the United States Navy Manpower and Personnel Systems – A Data Warehouse Approach

5. FUNDING NUMBERS

6. AUTHOR(S)
Conde, Douglas, J. Crownover, Cassandra A.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(maximum 200 words)*

Historically, stovepiped information systems have been developed to meet the needs of individual departments or users. Over time, attempts to increase the usefulness of these systems often involved adding layers of additional programming and data structures, resulting in complex and difficult to maintain legacy-based systems. The United States Navy enlisted personnel and manpower database system epitomizes this problem. The current system consists of several mainframe systems and a multitude of front-end systems that often require personnel managers to perform manual data extraction to execute routine activities. To illustrate the problem; we focus on Navy Enlisted Classification (NEC) reutilization, a critical aspect of the personnel assignment process. First, we present a series of contemporary database topics that form the basis for solving the problems associated with file-based legacy databases. Second, we provide details of the make-up and problems associated with the current system. Third, we develop a prototype relational data mart to prove the value of a data warehouse/data mart driven relational system. Fourth, using the prototype relational data mart as a source system, we use a contemporary OLAP application to prove the effectiveness of using a multidimensional data tool to analyze NEC reutilization. Finally, we discuss issues involving data quality and their impact on a data warehouse solution to integrating legacy systems.

14. SUBJECT TERMS
Database, Data Warehouse, Enterprise Data Warehouse, OLAP, Data Mart, Legacy system, Enlisted Manpower System, Enlisted Personnel System, Navy Enlisted Classification, NEC

15. NUMBER OF PAGES
132

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFI- CATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

THIS PAGE INTENTIONALLY LEFT BLANK

TOWARDS REENGINEERING THE UNITED STATES NAVY ENLISTED MANPOWER AND PERSONNEL SYSTEMS – A DATA WAREHOUSE APPROACH

Douglas J. Conde
LCDR, United States Coast Guard
B.A., State University of New York
College at Brockport, 1984

Cassandra A. Crownover
LCDR, United States Navy
B.S., United States Naval Academy, 1989

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
September 2000

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Historically, stovepiped information systems have been developed to meet the needs of individual departments or users. Over time, attempts to increase the usefulness of these systems often involved adding layers of additional programming and data structures, resulting in complex and difficult to maintain legacy-based systems. The United States Navy enlisted personnel and manpower database system epitomizes this problem. The current system consists of several mainframe systems and a multitude of front-end systems that often require personnel managers to perform manual data extraction to execute routine activities. To illustrate the problem; we focus on Navy Enlisted Classification (NEC) reutilization, a critical aspect of the personnel assignment process. First, we present a series of contemporary database topics that form the basis for solving the problems associated with file-based legacy databases. Second, we provide details of the make-up and problems associated with the current system. Third, we develop a prototype relational data mart to prove the value of a data warehouse/data mart driven relational system. Fourth, using the prototype relational data mart as a source system, we use a contemporary OLAP application to prove the effectiveness of using a multidimensional data tool to analyze NEC reutilization. Finally, we discuss issues involving data quality and their impact on a data warehouse solution to integrating legacy systems.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. BACKGROUND

Historically, programmers who translated detailed specifications into data structures and source code have created file-based legacy systems. Subsequently, these programmers changed careers and the attendant knowledge about the source code and data structure was lost. As additional system needs were identified, further complications emerged in the form of additional source code, front-end systems and complex program logic. The result of this process is typically islands of stove-piped systems that were not designed with integration and flexibility in mind. The current Navy enlisted personnel system for managing enlisted personnel data was developed in this manner. The current system consists of several mainframe systems and a multitude of front-end systems that often require personnel managers to perform manual data extraction from discrete database sources just to execute routine activities. This is inherently inefficient for users as well as sub-optimizing the amount of information that can be retrieved and integrated from these data sources.

The problem is clearly illustrated by Enlisted Community Managers (ECM's) and Enlisted Detailers who are presented with different tools and views of available and projected personnel inventory by various categories. ECM's utilize Mini-Stats augmented by the Navy Enlisted Classification (NEC) 1-to-5 Report generated by the Enlisted Distribution Projections System II (EDPROJ II) to project NEC inventories and reutilization rates. Detailers, on the other hand, rely upon additional systems such as the On-line Distribution Information System (ODIS), Enlisted Assignment Information System (EAIS) and the Readiness Information System (RIS) to provide actual available

personnel to assign to open requisitions via the Job Advertising and Selection System (JASS). The disparity between EDPROJ II personnel inventory numbers and those viewed as distributable by the enlisted detailers causes confusion and inaccurate application of these numbers when determining recommendations for individual personnel actions.

Enlisted personnel planners and managers need to effectively monitor and interpret substantial amounts of personnel information in the enlisted force. This information is maintained in several file-based legacy mainframe systems that were developed in the 1960's and 1970's. This thesis examines the file-based architecture used in the United States Navy enlisted manpower personnel systems and recommends a data warehouse driven approach for reengineering and integrating these systems. We document the major problems associated with the current system and detail the application of contemporary database technology to make enlisted personnel inventory data more useful and insightful to personnel planners.

B. OBJECTIVES

The primary question to be answered by this research is:

- How can a relational data warehouse or data mart and On-line Analytical Processing (OLAP) application unify Navy manpower legacy systems into an integrated database-driven system to provide ECM's and Enlisted Detailers with the proper views of required information, and the commensurate flexibility and scalability to make personnel management decisions effectively and efficiently?

Secondary questions to be answered by this research are:

- What are the deficiencies of the current file-based enlisted personnel database system?
- What are the features of contemporary database applications such as OLAP, data warehouse and data mart that can improve the current system?

- What performance metrics and associated “drill-down” dimensions are required for NEC reutilization? How can they be effectively implemented in a modern database system?

- What data quality issues need to be addressed in a migration plan to move from the current file-based system to implementation of an improved integrated enterprise approach (high level view)?

To highlight the benefits associated with modern database applications, we will create a relational database system containing data from various enlisted personnel data sources. Specifically, we will develop a prototype relational data mart that will help enlisted personnel planners track enlisted NEC reutilization. In addition, we will test and evaluate an OLAP application that will provide personnel managers with the capability to make personnel management decisions more effectively.

NEC reutilization was chosen because it provides a contemporaneous example at the cusp of enlisted personnel management. With the increased focus and dependence upon reutilization of NEC skills to improve readiness ashore and at sea, there is a commensurate need to track and project the status of NEC reutilization. Because of the recent redefinition of NEC reutilization to “having been in an NEC tour anytime in the last two tours or past six years”, current database systems are unable to supply this information readily. As a result, in some cases this process requires laborious manual data retrieval and manipulation each time this information is required. The prototype relational data mart and OLAP application will permit an ECM or Enlisted Detailer to produce reports required for making accurate personnel management decisions quickly and efficiently.

C. SCOPE AND METHODOLOGY

A search of contemporary literature concerning OLAP, data warehouses and data quality will be conducted. The current enlisted database system will then be used as a case study to understand the issues inherent to file-based legacy database systems. We will apply our research of database systems to the problem of enlisted NEC reutilization. This relatively innocuous problem provides a focused case study of problems associated with legacy systems, particularly their inflexible nature.

We will use the following systems analysis approach to develop the enlisted personnel relational database and OLAP prototype:

- Requirements Analysis - This phase involves interviews with users to determine the appropriate inventory strength performance metrics, their associated "drill down" dimensions, the source databases which must be accessed, the queries which are required, relevant business rules, and the reports and screen displays which the users wish to see in the application.

- Logical Database Design - Once the requirements gathering activities have been completed, a database design will be developed. This will include using semantic object modeling to represent the relevant objects and their relationships. The schema and object model will then be converted into a full relational schema implemented in Microsoft Access.

- Physical Database Design - The physical database design involves specification of appropriate primary and secondary keys, and implementation of all referential integrity constraints. This includes establishment of effective data quality standards for checking consistency of the input data, resolving missing values, and maintaining consistent levels of aggregation.

- Implementation and Testing - The database will be implemented in Microsoft Access.

- Database Analysis - Apply an OLAP application to the database to demonstrate analytical capabilities available in a contemporary database environment.

D. ORGANIZATION

This thesis is organized as follows. Chapter II presents a review of contemporary database technology. This chapter includes discussions on potential areas of database technology application within the Navy's enlisted management system.

Chapter III provides an overview of the current Navy enlisted database management system and the associated enlisted personnel Allocation, Placement and Assignment process. It provides insights on the problems associated with the current system and specifically discusses NEC reutilization.

Chapter IV provides a prototype relational data mart for the Navy enlisted personnel inventory system.

Chapter V demonstrates the ability of OLAP applications to "drill down" into a relational database. This will include implementation of an OLAP application as a management tool available to Navy enlisted personnel planners that can increase management effectiveness and efficiency.

Chapter VI addresses issues concerning data quality. This discussion will include a high level view of source system analysis, metadata, architecture, and management issues associated with moving from the current Navy enlisted file-based legacy database system to a contemporary data warehouse system.

Chapter VII provides a summary of our conclusions.

THIS PAGE INTENTIONALLY LEFT BLANK

II. DATABASE TECHNOLOGY REVIEW

A. INTRODUCTION

The United States Navy faces the daunting challenge of empowering knowledge workers with database systems that enable them to make effective and efficient strategic and tactical decisions. This is particularly important when a database system is used to make decisions that have enterprise wide significance and consequences. We consider database systems that meet these criteria to be called Critical Success Factor (CSF) systems. CSF's are conditions, events or results that must go well in order for organizations goals to be achieved and missions to be accomplished. In the vernacular, these database systems often involve "show stopper" issues that must be done correctly for an organization to be successful. A U. S. Navy database that we believe meets the CSF standard is the enlisted personnel database management system.

Note that we use the term "system" somewhat carefully when discussing "database systems." A system is defined in Merriam-Webster's Collegiate Dictionary as "A regularly interacting or independent group of items forming a unified whole" and "An organized integrated whole made up of diverse but interrelated and interdependent parts." However, as we will discuss in this thesis, the Navy enlisted personnel database "system" does not always reflect the "unified", "integrated" and "interacting" standards cited in the definition. We have found that enlisted personnel inventory and billet information is sometimes inaccessible to personnel planners or is maintained in a myriad of separate databases that does not provide easy or intuitive access. Of course, the U. S. Navy is not the only organization that finds itself in this situation. The heritage of disparate and

partitioned data spread among many file-based stove-piped systems is a common occurrence within both governmental and civilian organizations.

In this chapter we present a series of contemporary database management topics that can form the foundation for grasping and solving the problems associated with file-based legacy database systems. The topics include Data Warehouse, Data Mart, Online Analytical Processing (OLAP) and Data Mining. It is important to keep in mind that although the topics are discussed separately, there is a high level of inter-connectivity among them and the lines of separation are often blurred. For example, OLAP, data warehouse and data marts involve data integration, and data marts often depend upon a data warehouse. The business rationale is often similar for developing a data warehouse or deciding to use an OLAP or data mining application. For each topic, we conducted a literary search of current writings to provide a well-rounded review. Where appropriate, we have added amplifying information that specifically focuses on the issues being addressed in this thesis. A basic knowledge of this subject matter will be beneficial since this chapter is not intended to provide a complete and comprehensive explanation of each topic presented.

IT-21 provides minimum standards that will leverage IT solutions throughout the fleet. The priorities delineated in IT-21 need to be considered when entertaining new IT solutions. IT –21 provides the following guidance:

> The goal of IT-21 is to link all U.S. forces and eventually even our allies together in a network that enables voice, video and data transmissions from a single desktop PC, allowing war fighters to exchange information that is classified or unclassified, and tactical or non-tactical. To do this, systems must be built to **industry standards**, using **commercial off-the-shelf technology** (or COTS), **devoid of stovepipes**, in a client-server environment that allows the pull of just what information is needed in a way that's seamless to the user in the field. (Clemins, 1997)

Particularly meaningful when developing database-driven solutions is the following IT-21 guidance:

> **Relational databases** that can support web technology in accordance with the COE (such as Oracle, Sybase, SQL server and Access) will be used to support data requirements and application development. All process engineering initiatives that result in design and redesign of a **data collection and capture systems** must use COE compliant **relational database management systems (RDBMS)** software in order to ensure that RDBMS initiatives use COTS application software. (Clemins, 1997)

Specifically, IT-21 requires the following minimum standards be followed:

- Windows NT Server 4.0 is the Navy's standard network operating system.
- Windows NT 5.0 is the standard desktop operating system.
- Microsoft Office is the standard office suite software.

These facts make it prudent to compare the "openness" and capability of Microsoft products when developing IT solutions. Where applicable, each topic will include a short discussion of germane Microsoft products that can be considered when modernizing file-based legacy database systems. At no time should the discussion of any Microsoft or other company's product be considered an endorsement for the product.

B. REVIEW OF THE DATA WAREHOUSE AND DATA MART

The Information Technology (IT) community has adopted the term *data warehousing* to describe the requirements for, and the methods of collecting, consolidating, and storing enterprise-wide data within an organization. We begin our review discussing the Enterprise Data Warehouse (EDW) because we believe this strategy forms a solid foundation for developing an integrated and organized database system. Why do we need an EDW? One of the key factors governing an organization's success moving into the 21^{st} century will be its ability to disseminate useful information to decision makers throughout the organization. To achieve success, organizations must begin to provide true Business Intelligence (BI).

BI comprises systems that enable individuals to access information that describes an enterprise, to analyze it, to gain insight into its meaning, and to take action based on the results of the analyses via integration with other office functions. BI systems also present this information in an easy-to-digest form that facilitates effective decision-making. (Gartner Group, 2000)

A tremendous amount of research has been done to determine what characteristics an organization must have to be successful in the information age. A predominant theme is that shared knowledge must be in place to support a balance between decentralization and the speed and flexibility required to take advantage of unexpected opportunities and competitive changes.

If knowledge and intellectual capital are becoming key drivers of competitive advantage, then an intelligent organization is one that can broadly share its knowledge. Information systems play a role in creating and distributing that knowledge. Specifically,

the EDW, a central repository of subject-oriented data originating from an organization's transactions systems and external data sources, becomes a critical information system. The successful implementation of an EDW can have a significant effect in fostering a culture of knowledge sharing. (Gaskin, 1999)

1. Historical Perspective

To gain a historical perspective of the EDW, it is useful to look at a chronological history of the technology (Table 1). As is evident, the EDW is a relatively new technology that has progressed rapidly in order to meet market demands. The fast pace of evolution will likely continue in the future.

Period	Major Developments
Architectural Beginnings (1978-1988)	Studies at MIT target development of an optimal technical architecture. Digital Equipment Corporation builds a distributed network architecture and is the first to migrate to a relational database using RdB.
Enterprise Integration (1988)	IBM introduces the term "Information Warehouse" to tackle the problem of enterprise integration. However, the architecture is still not viable.
Enterprise Data Warehouse (1991)	Inmon publishes "Building the Data Warehouse" that focuses on how to build an EDW.
EDW Divergence (1996-1997)	In a bid to be first to market, many companies promote different architectural versions of the EDW, for example, relational architectures, multi-dimensional cubes, data warehouses, data marts and operational data stores.
EDW Synthesis (1998)	The enterprise solution approach takes precedence in most products. The EDW is most often viewed as overall application architecture and not added as an afterthought.

Table 1. EDW Chronology (Haiston, 1999)

2. Data Warehouse Definitions

Generically, an EDW contains data captured from one or more operational systems. The data is transformed, cleansed, integrated and loaded into a separate subject-oriented EDW. It contains detailed (and possibly summarized) data recording business operations over a period of time that may range from a few months to many years. The historical nature of the data supports detailed analysis of business trends over a period of time. The EDW is used for making both tactical and strategic business decisions

covering multiple business areas. Bill Inmon, a well-known EDW expert defines an EDW in a technical way:

- **Subject-oriented**: Mandates a cross-functional slice of data drawn from multiple sources to support a diversity of needs.

- **Integrated**: The process of mapping dissimilar codes to a common base, developing consistent data element presentations and delivering this standardized data as broadly as possible.

- **Time-variant**: Calls for storage of multiple copies of the underlying detail in aggregations of differing periodicity and/or time frames. You might have detail for seven years along with weekly, monthly and quarterly aggregates of differing duration. The time variant strategy is essential, not only for performance, but also for maintaining the consistency of reported summaries across departments and over time.

- **Nonvolatile**: Once a row in a table is written, it is never modified. This is necessary to preserve incremental net change history. This, in turn, is required to represent data as of any point in time. (Inmon, 1999)

One of the principal reasons for developing an EDW is to integrate operational data from multiple sources into a single and consistent architecture that supports analysis and decision-making within the enterprise. The primary objective of the EDW is to "support analysis of data for business needs, which stands in sharp contrast to the traditional use of database technology for capturing operational data." (Dolk, 1999)

An EDW is typically a blending of technologies, including relational and multidimensional databases, client/server architecture, extraction/transformation programs, graphical user interfaces, and more. An EDW turns raw information into a useful analytical tool for decision-making. Raw information is normally kept in online transaction processing (OLTP) systems, which track day-to-day operations. However, OLTP systems aren't well suited for answering questions that affect the past, present and future directions for an organization. To answer those kinds of questions, a company

needs an analysis system with the ability to perform ad hoc queries and create specialized reports.

3. Star Schema

EDW's are designed somewhat differently from "conventional" databases, using star schemas rather than standard Entity-Relationship (E-R) diagrams. E-R diagrams contain entities such as Students and Courses. These entities have relationships between them such as Students enrollment in Courses. The basic premise of star schemas is that information can be classified into two groups: facts and dimensions. Facts are the core data element being analyzed whereas dimensions are attributes about the facts. Most, if not all, analysis is based on these dimensions, hence the term dimensional analysis. The star schema is built for simplicity and speed. The assumption behind it is that the database is static with no updates being performed online. Dimensions store attributes for values in the fact table. Denormalization of a star schema has the direct effect of minimizing the number of joins, therefore increasing performance (one fact table references numerous dimension tables). Join processing is accelerated by indexes on fact table columns used as foreign key references to the dimension table's primary keys.

Visually, a star schema consists of a hub and spokes where the hub is a fact or performance metric and the spokes are dimensions that are relevant to a metric. The star schema design is based on direct paths (joins) between the hub and spokes. It's also significant that, in the typical star schema, the fact table is much larger than any of its dimension tables. This becomes important when considering performance issues associated with star schemas. See Chapter V, Section C for an example of a star schema.

4. Data Warehouse Scalability

> Scalability is the ability to process a given workload faster by simply adding resources. Furthermore, scalability also implies that performance should be consistent and predictable as new resources are added. (Sun Microsystems, 1998)

We include scalability as a separate topic because we feel it has traditionally not been given the high level of attention it deserves during system development. Almost without exception, the volume of data a database system must handle grows faster than forecasted. For an integrated database-driven enlisted manpower system to be successful, it must be highly scalable. A fundamental question must be answered, how can data warehousing scale to meet the need of an increasing volume of data? What technologies can be employed to meet the problem?

A study by Sun Microsystems Inc. provided insight into some of the technical issues concerning the scalability of data warehouses. The study indicates that scalability is possible using large-scale server technology called Symmetric Multiprocessing Systems (SMP) that coincide with the use of parallel database systems. The study states that fully parallel processing allows nearly every aspect of data warehousing to run more quickly on SMP hardware and significantly improves performance. We do not attempt to give full coverage of the study here, but the results of an enterprise level test were impressive and noted in Table 2 below.

Hardware	Time to Completion
Sequential 4-processor	32 hours, 34 minutes
Parallel 4-processor	9 hours, 54 minutes
Parallel 8-processor	4 hours, 27 minutes
Parallel 12-processor	3 hours, 1 minute

Table 2. Elasped Time From Tests (Sun Microsystems)

5. Why Do Data Warehouses Fail?

Throughout our literature search, we discovered a recurring set of reasons why data warehouses fail to meet business requirements, or become outright failures. The lessons learned from these failures provide valuable insights to consider prior to developing a data warehouse. The lessons learned are summarized below:

- Failure to define the business purpose of the data warehouse, i.e., what well-defined business problem(s) will the data warehouse solve?
- Failure to account for scalability issues that lead to poor performance.
- Population of the data warehouse with data that has not been properly refined and cleansed.
- Selection of the wrong architecture.
- Implementation of a data warehouse as a huge top-down effort with little input from users.
- Development of stove-piped data marts that did not integrate across the enterprise.

6. The Data Mart

A data mart can provide many of the same benefits associated with an EDW, i.e., a data source with data that has been cleansed and transformed and provides the ability to perform complex analysis of the data. A data mart contains a subset of data that is of value to a specific business unit, department, or set of users. This data subset may be captured from one or more operational systems. Data marts usually contain summarized historical data for a specific business area of an organization. The attraction of a data mart is quicker development and lower cost than an EDW. The cost of large EDW projects can scale into the millions of dollars and take a year or more to implement. Data marts built directly from operational systems can often be developed in a matter of

months at a significantly lower price. However, the low cost of data mart development can lead to problems as each business area builds its own independent data mart.

7. Microsoft Data Warehouse Products

Microsoft has taken a comprehensive approach to data warehousing in planning their SQL Server 7.0 database product. The following components are included in SQL Server 7.0:

- **Integrated OLAP Server**: Using the online analytical processing (OLAP) services, it is possible to develop a variety of enterprise solutions, such as corporate reporting and analysis as well as data modeling and decision support. With support for full multidimensional, relational, and hybrid OLAP implementations, you can choose the data model that best suits your application. OLAP Services also offers intelligent aggregations, resulting in smaller databases, improved performance, and shorter initial and incremental load times. See Section C of this Chapter.

- **Data Transformation Services**: Data Transformation Services (DTS) makes it easy to import, export, and transform heterogeneous data using OLE DB, Open Database Connectivity (ODBC), or text-only files. DTS also eliminates the need for user intervention by allowing you to import or transform data automatically on a regularly scheduled basis.

- **Microsoft English Query**: With English Query, you can create applications that accept natural language queries (questions written in plain English) instead of complex SQL queries.

- **Repository Integration**: SQL Server 7.0 includes Microsoft Repository and the Open Information Model, which together help to integrate and share metadata about SQL Server databases, OLAP Services, Data Transformation Services, and English Query. See Chapter VI, Section E-1.
(Microsoft Corp, 2000)

C. REVIEW OF ONLINE ANALYTICAL PROCESSING (OLAP)

Organizations can prosper or fail according to the sophistication and speed of their information systems, and their ability to analyze and synthesize information using those systems. The numbers of individuals within an enterprise who have a need to perform more sophisticated analysis is growing. A contemporary database analysis tool that enables knowledge workers to analyze enterprise wide issues are OLAP applications. OLAP applications allow knowledge workers to manipulate data quickly, intuitively, and flexibly using familiar business terms in order to provide analytical insight. More specifically, OLAP is a category of software technology that enables knowledge workers to gain insight into data through a wide variety of possible views of information that has been transformed from raw data. The OLAP Council, a group dedicated to OLAP development, defines OLAP as:

> A category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information. OLAP transforms raw data to reflect the real dimensionality of the enterprise as understood by the user. (OLAP Council, 1997)

Many consider OLAP's ability to view data from a multidimensional perspective as its most important attribute. Multidimensional analysis allows end users to gain insight into the meaning of data contained within databases. It can provide insights, trends and solutions that are otherwise very difficult to glean from the data. A multidimensional view enables a user to "slice and dice" data, thus providing a user with the ability to analyze data across many dimensions and levels of aggregation. OLAP can also utilize client-server architecture and networks to consolidate and control

information, while making the information accessible to those that need it via web browsers.

OLAP applications depend on a user to input a question or hypothesis. It enables an analyst to ask speculative "what-if" and "why" scenarios executed within the context of a historical perspective. For example, a typical OLAP question might be, "why do enlisted personnel in aviation billets have a higher NEC reutilization than enlisted personnel in supply billets and how will planned changes impact the future?" This question-directed format requires a user to have significant business knowledge in order to ask meaningful questions. It also reveals how OLAP can be used to combine analysis of historical data with future projections.

OLAP functionality is characterized by the following activities:

- Calculations and modeling applied across dimensions, through hierarchies and/or across members.
- Trend analysis over sequential time periods.
- Slicing subsets for on-screen viewing.
- Drill-down to deeper levels of consolidation.
- Reach-through to underlying detail data.
- Rotation to new dimensional comparisons in the viewing area.

1. Evaluating OLAP

Dr. E. F. Codd, father of the relational database, provides one of the original definitions of OLAP. The definition includes twelve technical evaluation criteria for evaluating an OLAP application that can in turn be used to form the basis for evaluating the architectural framework of an OLAP application.

The twelve rules for evaluating an OLAP application are:

- Multidimensional Conceptual View
- Transparency
- Accessibility
- Consistent Reporting Performance
- Client-Server Architecture
- Generic Dimensionality
- Dynamic Sparse Matrix Handling
- Multi-User Support
- Unrestricted Cross-dimensional Operations
- Intuitive Data Manipulation
- Flexible Reporting
- Unlimited Dimensions and Aggregation Levels

Critics claim the Codd rules are lacking because they are unsuitable for determining OLAP compliance. As a result the OLAP Report (Nigel Pendse and Richard Creeth) has produced its own definition based on what it calls the FASMI (Fast Analysis of Shared Multidimensional Information). Table 3 lists the meaning of each of these terms.

Term	Definition
Fast	Means the system is targeted to deliver most responses within around five seconds, with the simplest analyses taking no more than a second and very few taking more than 20 seconds.
Analysis	Means the system can cope with any business logic and statistical analysis that is relevant to the application and the user, and keep it easy enough for the target user.
Shared	Means the system implements all the security requirements for confidentiality and - if multiple write access is needed - concurrent update locking at an appropriate level.
Multidimensional	The system must provide a multidimensional conceptual view of the data, including full support for hierarchies and multiple hierarchies, as this is certainly the most logical way to analyze businesses and organizations.
Information	All of the data and derived information needed, wherever it is and however much is relevant for the application.

Table 3. FASMI Definitions (OLAP Report, 2000)

2. OLAP Approaches

OLAP vendors have typically used two approaches in their products. The first approach is to use corporate data stores to build multi-dimensional databases, or cubes, specifically for the purposes of OLAP. This is typically called Multidimensional OLAP (MOLAP). The second approach is to build a traditional relational database and construct a series of OLAP queries and applications to directly access the data. This is typically called Relational OLAP (ROLAP). OLAP technology has been continuously

progressing over the last few years. This has caused a lot of confusion as the various OLAP camps tout their design as best. Some of the newest OLAP products contain elements of MOLAP and ROLAP, and are called Hybrid OLAP (HOLAP). Finally, Desktop OLAP (DOLAP) applications are client-based OLAP products that are easy to deploy and have a low cost per seat. They typically have good database links, often to both relational and multidimensional servers. DOLAP applications usually have limited functionality and capacity compared to the more specialized OLAP products. The rapid evolution of OLAP technology means IT managers must carefully evaluate their business requirements prior to selecting an OLAP solution. Additional information for MOLAP and ROLAP is offered below since virtually all OLAP applications use one of these architectures.

MOLAP uses organizational data to build a series of multi-dimensional cubes of data specifically designed to address a specific or narrow range of inquiries. The data is typically aggregated or pre-calculated. This allows OLAP queries to be fast, since the calculation of summary data is already done. For example, enlisted Detailers requiring NEC reutilization data for all aviation rates in a particular region would be provided an aggregated multidimensional cube with the required data. Typically there is no method of navigating around the cube to the raw data; therefore, the users' access to raw data is limited. Because of its excellent performance capabilities from a user perspective, MOLAP is the single most widely used approach in OLAP.

ROLAP products adapt traditional relational databases to support OLAP. The term ROLAP implies that an OLAP server accesses a relational database rather than a multidimensional database. An EDW is typically used as the data source. The OLAP

engine performs data recovery by generating SQL to retrieve data and then presenting the output to a user. Because data is accessed directly from within the EDW, ROLAP systems do not have some of the pre-processing and aggregation issues experienced with MOLAP systems. However, depending upon the analysis conducted and the resulting query executed, ROLAP products have a risk of encountering performance problems.

3. Data Depth and Data Breadth

An interesting issue surfaces around data depth and data breadth. Data depth refers to the level of data granularity, whereas data breadth refers to the number of dimensions and attributes that can be analyzed. (Hurwitz Group, 1998)

Naturally, different levels of users will have varying needs for data depth and data breadth. However, all personnel require information at a depth and breadth necessary to analyze problems and provide new opportunities. Higher-level personnel management would benefit from enlisted data that has been summarized in reports that provide a broad Navy-wide perspective. For example, they might want to compare NEC reutilization across all enlisted rates. Lower-level operational users would typically require access to data aggregated at a lower level of granularity.

A Navy enlisted manpower database would have dozens of dimensions and attributes such as such as a Time dimension (months, years), Geography dimension (cities, countries) Rate dimension (electrician, aviation mechanic) and Billet dimension (rate, NEC). These are needed to establish a complete picture of personnel and billet data when making human resource decisions. However, each individual personnel planner might require aggregation only for the specific enlisted rate(s) they are responsible for.

The ability to aggregate information that meets individual requirements from a singular data source is a substantial benefit of OLAP.

4. OLAP Metrics

In order to provide additional perspective on the effectiveness of the various OLAP architectures discussed, Table 4 presents performance metrics we developed to rate each OLAP architectural approach. HOLAP was not included since it has elements of both MOLAP and ROLAP. Our conclusions are not based on direct experimentation of various products and architectures, but solely upon our literature search and limited testing performed for this thesis. The metrics chosen are those regularly identified as critical operational performance measures during our literature search. The rating is on a 1 to 5 scale, with 1 being the lowest score and 5 being the highest. Obviously, OLAP performance is highly dependent on many factors such as hardware, number of users, network speed, etc.; however, we believe the results provide a good general survey of the strengths and weaknesses of each OLAP architecture.

OLAP Metric	MOLAP	ROLAP	DOLAP
Scalability	4	4	3
Support for large atomic data sets	4	5	3
Ad-hoc query capability	3	5	3
Number of Dimensions	3	4	3
Granularity Level	4	5	4
Effect of Data Volatility	3	5	2
Open architecture standards	2	4	3
Cost	3	2	5
Speed	5	3	5
Third party Support	4	2	4
User Friendliness	4	3	5
Ease of deployment	4	3	5
Complex statistical analysis	4	3	2
Batch processing time	3	5	3
Installed user base	4	2	5

Table 4. OLAP Performance Metrics

5. OLAP Products

It is very important for each IT manager to evaluate their current and future business requirements prior to deciding which type of OLAP product will provide the best solution. Similar products can produce disappointingly opposite results for different organizations. In the end, business requirements must drive the selection of the proper architecture and product. We feel that an integrated database-driven enlisted manpower system would be best served by a MOLAP or HOLAP application. Table 5 lists many of the commercial off-the-shelf OLAP software packages currently available and the architecture used for the product.

Software Package	MOLAP	ROLAP	HOLAP	DOLAP
Hyperion Essbase	X			
PowerPlay Enterprise Server	X			
Hyperion Enterprise	X			
SAS CFO Vision	X			
Comshare FDC	X			
Microstrategy		X		
IBM OLAP Server		X		
Informix Metacube		X		
Oracle Discoverer		X		
Microsoft OLAP Services			X	
Pilot Analysis Server			X	
Seagate Holos			X	
Oracle Express			X	
Brio Enterprise				X
Business Objects				X
Cognos PowerPlay				X
Personal Express				X

Table 5. OLAP Software Packages

6. Microsoft OLAP Products

Microsoft offers a data warehouse/OLAP business solution. It is implemented with the following commercial off-the-shelf products:

- Microsoft SQL Server 7.0
- Microsoft OLAP Services (built into SQL Server)
- Microsoft Excel 2000

Improved data warehousing support is a major part of SQL Server 7.0. The biggest new data warehousing-related feature is the addition of Microsoft's new OLAP server, known as the Microsoft SQL Server OLAP Services. Microsoft SQL Server OLAP Services supports Multidimensional (MOLAP), Relational (ROLAP) or Hybrid (HOLAP) implementations.

Microsoft SQL Server OLAP Services integrates with Microsoft Office applications such as Excel Pivot Tables and includes the Object Linking and Embedding

Database (OLE DB) for OLAP provider and ActiveX Data Object Multidimensional (ADOMD) that enable custom access. One of SQL Server 7.0's new features is the addition of the Microsoft Management Console (MMC) for SQL Server administration. Figure 1 presents the MMC view of Microsoft SQL Server OLAP Services and Figure 2 presents the overall architecture of the Microsoft solution.

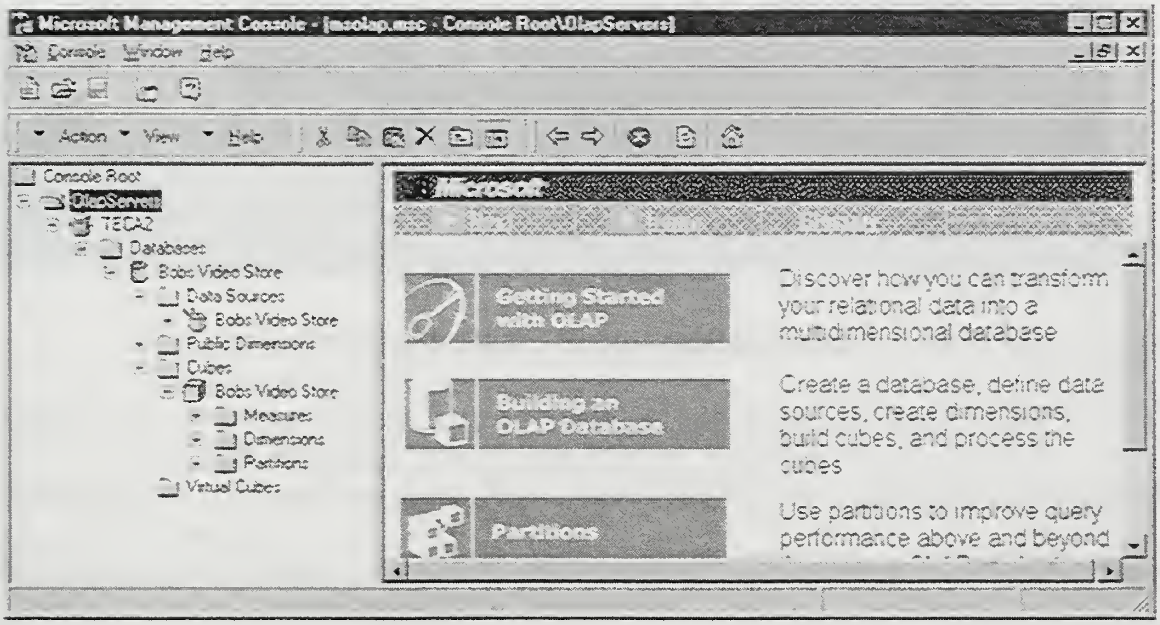


Figure 1. Microsoft SQL Server OLAP Services (Microsoft Corp.)

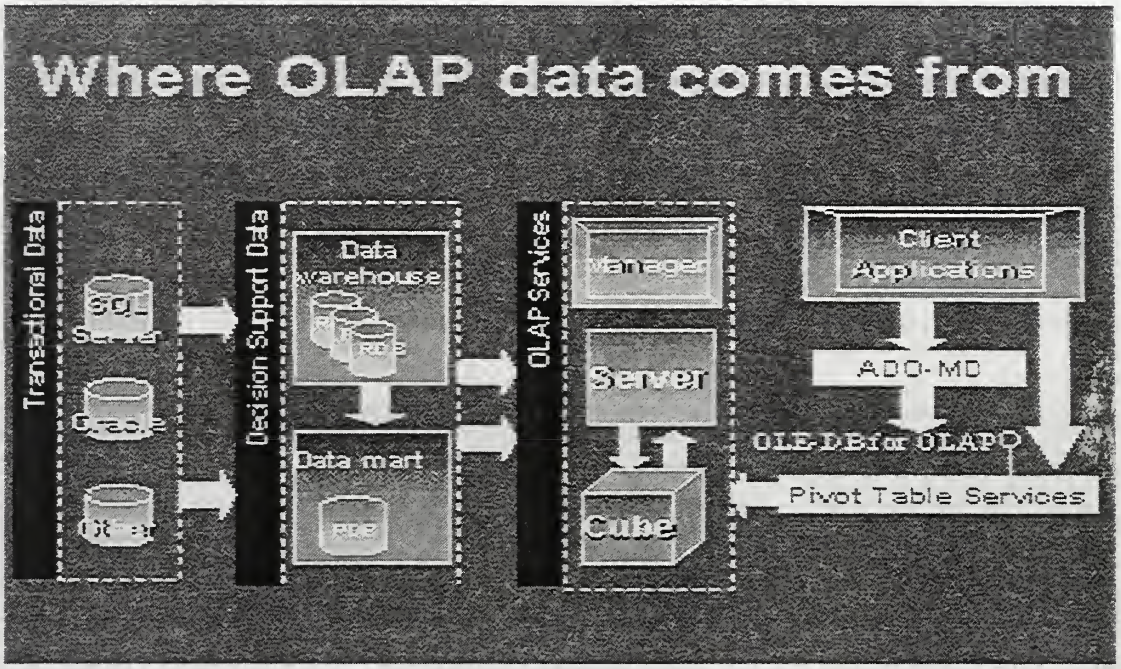


Figure 2. Microsoft HOLAP Architecture (Microsoft Corp.)

This solution pulls data from operational data sources into a relational Data Warehouse or Data Mart. OLAP Services is used to create and manage multidimensional cubes (MOLAP) or generate SQL to retrieve data (ROLAP). These implementations use open standard ODBC or OLEBC.

Microsoft's OLAP solution requires that separate compatible software be purchased to use it since OLAP Services uses a query language different from the one SQL Server uses (Excel 2000 is one client, but others are available from companies like Cognos Corp. and Seagate Software).

7. OLAP Hardware

OLAP involves more than the implementation of software. When developing an OLAP solution, the hardware platform is often a critical element of success or failure. The following list offers general guidance on hardware considerations when developing

OLAP with Microsoft SQL Server and OLAP Services. Many of these suggestions will also improve performance for all OLAP applications.

Memory:

- More memory is better; a larger cache helps process larger dimensions and increases query performance.
- A large memory space is needed to process tens of thousands of members in dimensions in large systems.
- Buffering can dramatically increase processing time.

Disk Speed:

- ROLAP and HOLAP read large amounts of data at a low level of granularity; faster disk speeds will improve performance.
- MOLAP performance is enhanced when a lot of data is cached in memory and read directly out of memory.
- MOLAP – Maximize I/O by using multiple drives in a RAID array.

CPU Speed:

- A single CPU is adequate for a few users.
- Querying is inherently a parallel process; to increase performance, take advantage of multiprocessing.
- Process cubes in parallel; use Decision Support Objects that come with cubes.

THIS PAGE INTENTIONALLY LEFT BLANK

III. CURRENT ENLISTED SYSTEM

A. INTRODUCTION

Enlisted personnel planners need to effectively monitor and plan for the optimal reutilization of skills within the Navy. The detailing system that is in place, including all of the programs and databases, is very complex, and it is difficult to ensure that skills are being reutilized in such a way that is cost effective and manageable.

Data on enlisted personnel is collected through an intricate network of programs and interfaces. Through this system data is collected, validated and finally processed by the Navy Enlisted System (NES). The NES in turn produces a file called the Enlisted Master File (EMF), which is a flat file containing data on each enlisted person in the Navy. This file is used by many systems to create queries, reports and planning data for enlisted detailers and community managers. These tools are then used to detail and plan enlisted manning now and in the future. The EMF will be the principal source of data for this project. It holds current and historical data on each Sailor's skills, assignment history and a plethora of additional personal data.

B. NAVY ENLISTED CLASSIFICATION CODES

Navy Enlisted Classification (NEC) Codes supplement the Navy rating system in identifying individual skills, knowledge and qualifications that are not rating wide. NEC's are necessary to manage specialized billets and personnel to aid in the efficient detailing and projection of Navy enlisted personnel. Billets that require specific skill sets are coded in such a way that only personnel that possess those skills should be assigned. This ensures that once a person receives specialized training, that may be costly, they are then used to meet the needs of the Navy in that field.

NEC reutilization occurs when a person that possesses a specific NEC and has used it within a particular timeframe is ordered to a billet that requires that NEC. The timeframe currently being used is six years or within the previous two commands. Measurements are currently being made by Enlisted Personnel Management Center (EPMAC) utilizing the EMF. The data is extracted using the RIS system, imported into a spreadsheet and manipulated to produce navy wide and community wide reutilization rates. These results are quickly outdated, difficult to understand, and omit important information.

There are other problems associated with the current system of tracking NEC reutilization. For example, there is no system in place to track reutilization for a particular command or to project reutilization in the future. There is no system that tracks where personnel who meet the NEC reutilization criteria are being ordered to in lieu of a billet that requires their particular skills. There are many rules that impact the calculations of NEC reutilization and some that are not very easy to model. There are pairs of NECs that indicate the same skill, but are different numerically for shore and sea billets. There are pairs that indicate the same skill and are different numerically for technician and supervisor. If this data could be included in the calculation the reutilization rates may increase significantly. When calculating reutilization it would be useful if this information could be contained in the calculation of the data. Unfortunately, this is near impossible with the tools that are being utilized today.

While calculating reutilization is difficult, the actual detailing of enlisted personnel with specific skills is an extremely cumbersome task. Personnel with extensive skills and costly school backgrounds are routinely detailed to billets that do not require

those skills. People that do not possess the skills are detailed to positions that require them, at a significant cost in additional training. To understand the problem, and why personnel with NEC qualifications are ordered to billets with no NEC requirement, a discussion of the detailing process is required.

C. TRIAD OF DETAILING

The Navy relies upon the triad of detailing to distribute personnel throughout the fleet. The three functions making up the triad are Allocation, Placement and Assignment.

Allocation is broken into two main functions as shown in Figure 3. Primary responsibility for Allocation lies with PERS-45. EPMAC is responsible for allocation of all personnel E-3 and below without a rate (Gendets). PERS-45 responsibilities include developing projections of distributable strength and distribution among the Manning Control Authorities (MCAs) in order to maintain a balance throughout the fleet. The four MCA's are CINCLANTFLT, CINPACFLT, BUPERS and Commander Naval Reserve Force (CNRF). Based on information collected from each MCA, projections are made on what the Navy's distributable strength will be 7-9 months in the future. After projections are made, planners use this information to determine how personnel will be allocated across the MCA - composite groups. These groups are Air, Surface/Subsurface and Shore Activities. The output of the allocation process is a report depicting how many sailors from each rating/paygrade must be sent to each MCA/composite group.

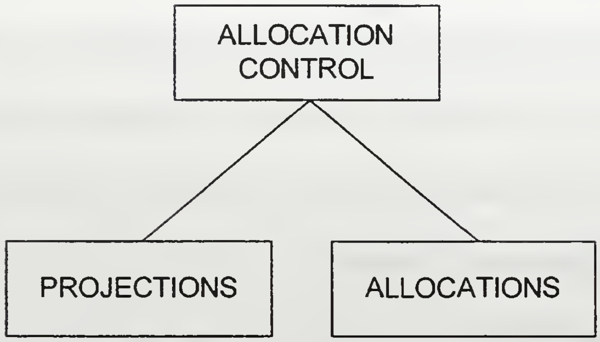


Figure 3. Allocation Breakdown

The second leg of the Triad, Placement, is accomplished by the four MCA's and EPMAC. It is broken down into Manning Control Authority and Placement, as shown in Figure 4. With the information provided from the allocation process, the MCA's determine the necessary manning of each ship, squadron and shore command. They further break the requirements into paygrade, rating and NECs. Planners at the MCA's compare projected allocation numbers with the number of Billets Authorized (BA) at each activity and develop a plan that ensures fair share distribution of the personnel assets. From this planning the Navy Manning Plan (NMP) is developed which reflects the combined manpower requirements of all MCA's and Navy commands and "fair share" guidance. EPMAC coordinates and tracks the manning needs of the MCA's and is responsible for the development of the NMP. The final product of this leg is the Requisition Report, produced by EPMAC, which includes all of the billets that need to be filled in the seven to nine month window, and the priorities assigned by the MCA. This report is the primary tool used to communicate an activity's personnel needs to the Detailers.

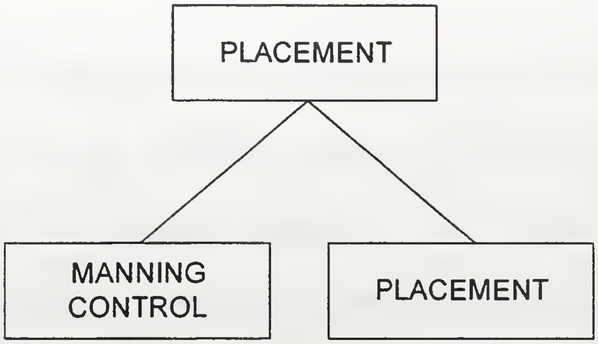


Figure 4. Placement Breakdown

Assignment is the last leg of the triad and is the actual detailing of personnel. This is accomplished by PERS-40 for designated strikers and Petty Officers, EPMAC for Gendets, and NRPC for TAR enlisted. Assignment is accomplished by matching supply and demand as illustrated in Figure 5.

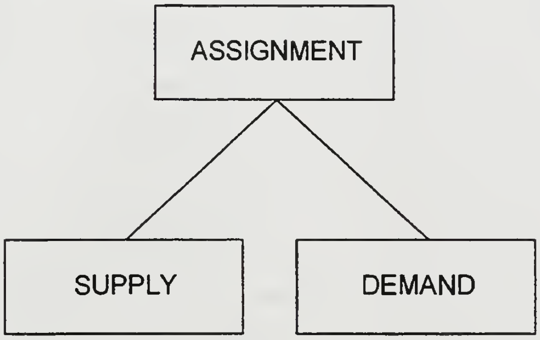


Figure 5. Assignment Breakdown

The goal of the detailers is to balance the supply of sailors with the demand of billets by matching those available for orders with requisitions on the Requisition Report. The Requisition Report, which is updated bi- weekly by EPMAC, is the primary source for assessing the demand for sailors. The supply side is more complicated because Detailers must carefully evaluate each individual who is available for orders and research their previous assignments, training, NEC history, performance, and duty preferences. The outcome of the assignment phase is a set of orders.

D. RESOURCES

The set of database systems that make up the total distribution system are referred to as the Navy Military Distribution System (NMPDS). Each part or subsystem was designed to meet a specific need or compile specific facts related to Allocation, Placement or Assignment. The Distribution Management Support System (DMSS) is an umbrella system that includes a number of different programs in support of either personnel inventory or manning readiness. Figure 6 depicts the interaction between the major systems that make up the enlisted distribution system.

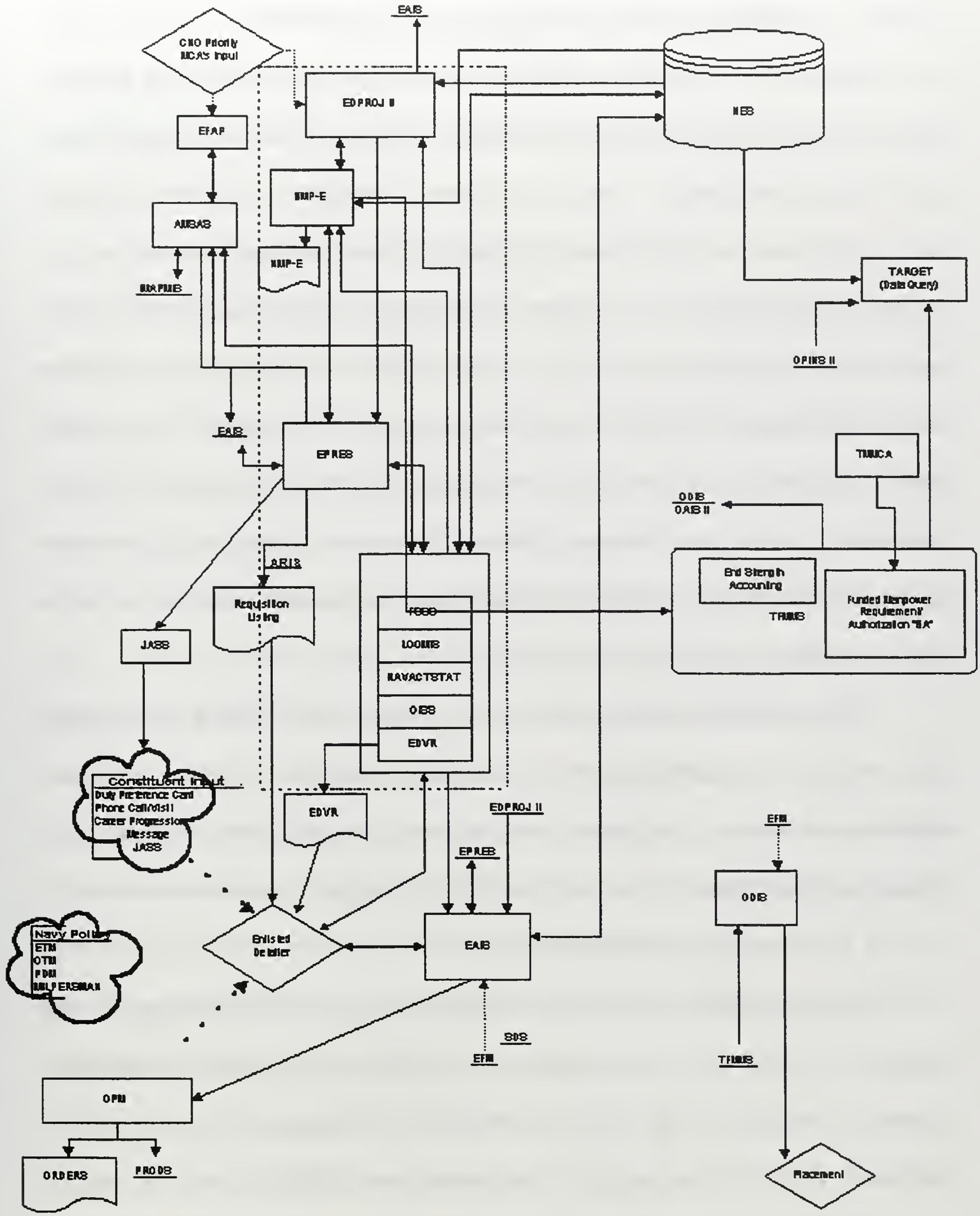


Figure 6. Navy Enlisted Distribution System

1. Enlisted Distribution Projection System (EDPROJ)

EDPROJ is a mainframe COBOL based system that is resident at DMC in Chambersburg, PA and maintained by EPMAC. EDPROJ runs at the end of every month, using approximately 12 hours of mainframe operating time. It receives inputs from the EMF and the billet database (TFMMS). It measures those personnel who are currently attached to an activity and those that expected to be attached in nine months. Personnel not attached or under orders are further broken down and the number of those that are distributable are divided equally between sea and shore billets. Using these inputs, it projects where personnel should be assigned based on a) NECs, b) CNO priority billets and c) fair share of remaining billets. The result is the Enlisted Distribution Allocation Report (EDAR), which shows the number that will be allocated to each of the MCA's, broken down by composite groups (air, sea and shore).

Other products include an Allocation Tracking Module (ATM) that provides online tracking of allocation measures for the detailer. EDPROJ is the only projection system that measures distributable strength using current billet and personnel information. The information is not real time and lags changes in personnel availability.

2. Navy Manning Plan (NMP)

NMP is a COBOL based mainframe application located in Chambersburg, PA and maintained by EPMAC. Since personnel assets rarely equal manpower requirements, distribution managers use the NMP to determine how shortages and excesses shall be distributed. The NMP runs monthly in conjunction with EDPROJ, taking information from the EDPROJ system and processing it down to the activity level. It takes input from the MCA's, the EMF and TFMMS and spreads the distributable strength across the

various activities. It provides requirements of MCAs and fair share distribution guidance of billets. Using the projected level of assets for a composite and the billets authorized for each activity, the NMP determines the most equitable level of manning an activity can expect for each detailing composite.

3. Enlisted Personnel Requisition System (EPRES)

EPRES is a COBOL based mainframe application located in Chambersburg, PA and maintained by EPMAC. EPRES is used to track projected gains and losses and identify billets that need to be filled in the seven to nine month window. This is the source for Requisition Reports that are produced twice monthly by EPMAC. EPRES measures the personnel needs of each activity for the forthcoming nine months by comparing projected onboard assets to the NMP and NMP to billets authorized. The result of this measurement is a personnel requisition indicating the rating or NEC required by the activity and the month in which the individual is required onboard.

4. Enlisted Assignment Information System (EAIS)

EAIS is the system that is used by detailers to research and assign personnel to billets. It contains the Requisition Posting Module that contains billet vacancy information, personnel information available by SSN, and an order-writing module. Information that is input by the detailers for order- writing is fed back into the EMF to update to the record.

5. Readiness Information System (RIS)

RIS includes personnel, billet and NMP information. It uses data that is created and maintained in a batch mode and made available through online files. The files are separated into several programs, which are accessed via data entry and menus. Data can be manipulated within each program to create reports. Each user must possess extensive knowledge of the system to use it effectively. The manual for the system is over 600 pages long.

6. Online Distribution Information System (ODIS)

ODIS provides detailers with the ability to conduct ad-hoc queries of personnel and activity bases in support of distribution processes. Detailers use this system in conjunction with EAIS to manage the detailing of their personnel.

7. NES Data Dictionary

BUPERS has created a data dictionary, which identifies and defines each field within the EMF. This is useful when extracting data in flat files from the EMF. It is also useful in ensuring that new systems maintain a baseline of integrity for those fields defined. An extract of the Data Dictionary is shown in Figure 7. It displays the field name, position, length, description, and source of data for one data element. There are 110 fields in the EMF.

Data Element Full Name: SOCIAL SECURITY NUMBER

EMF Copylib Name: SSN

EMF Position: 0001-0009

EMF Length: 9

NES TAC: 1xx, 2xx, 344, CAC, MOB, QC0

Description: The unique identification number assigned by the Social Security Administration (SSA).

Valid Values: Valid ranges issued by the SSA.

Auth Sub Sources: PERS-1023C (OCARS, OPINS), PERS-1024E, PERS-103 (SDS), PERS-29

(COMPASS), PERS-47, PERS-842, EPMAC, MEPCOM, NRPC

Comments: References: (a) MILPERSMAN 4610100; (b) CRUITMAN-ENL 1-I-13, 8-I-2, 8-I-17; and

© PAYPERSMAN 90658. This data element complies with the DOD standard data element SO-CA-AB in the Manual for Standard Data Elements, DOD 5000.12-M.

Figure 7. Data Dictionary Extract

E. KNOWN ISSUES

The NES has been the Navy's enlisted personnel system since 1973. There have been many different organizations involved in developing and maintaining these systems throughout the years. There is very little knowledge concerning the many lines of code for most of these systems, as the personnel involved in the development have changed numerous times. The NES is made up of at least 260 modules, 900 programs, at least

400,000 lines of code, and interfaces with at least 25 other programs. NES runs on a periodic basis, collecting and updating data from many different systems. It is complicated and may access a particular system more than once, each time collecting from different fields.

ODIS has proven burdensome and complicated to many. Some detailers are adept at pulling the information from the system that is needed, while others are forced to manually look up all information on a case-by-case basis. It assumes, in some cases, an individual's ability to create and manage SQL queries. In practice, the query system is used by some and not by others.

EAIS requires a detailer to move record by record through individual records of available personnel to determine who meets the requirements of the job that is being assigned. This system is both time-consuming and inaccurate, as it depends on the individual's thoroughness to ensure a proper fit for the personnel and gaining command. Many times the qualifications of the person are overlooked when detailing to positions that don't require certain qualifications, which can lead to inefficient detailing of those personnel and the underutilization of certain NECs.

NEC's are a commodity in the Navy. Effectively reutilizing them can save a great deal of money in training and retraining. EPMAC is able to provide NEC reutilization numbers on a quarterly basis. These numbers are derived from the EMF using DNEC information for the current and previous two commands. The numbers are calculated based on the total number of personnel in the Navy that have been detailed. The problem is that this information is not available for planning purposes, and it is not broken down by command. ECMs receive complaints from individual commands, yet they are unable

to easily provide valid numbers of personnel sent to each command with required skills. They are forced to manipulate data using existing systems and manual updates. There is no tool to provide them with information on reutilization by billet and command.

Currently managers use the tools that are available to them to import data into Excel™ files and manipulate them manually to determine reutilization. This sometimes is done by looking up data on individual sailors, which is time consuming. Information is sometimes not accurate in the system, so the person computing the reutilization rates will change the numbers based on knowledge they possess about an individual, or what they interpret from a record.

The current systems do not provide the ECM the ability to make real time decisions and projections. They must rely upon outdated information and outdated systems to provide information. There is no tool for the ECM to make daily decisions and projections for the assignment of sailors.

F. FUTURE

The personnel systems that the Navy maintains are large, burdensome, and difficult to use. Although the systems are old, they store data in flat files, which can be imported into database software. We can use the information that resides in these systems to create management tools to view and summarize the data that exists. The imported flat files or spreadsheets will populate relational or object oriented databases, which can then be used to query necessary information.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. RELATIONAL DATA MART PROTOTYPE

In order to provide a useful planning tool, the available data must be stored in a format that will facilitate easy retrieval and query reporting. To do this we will create a prototype relational data warehouse with predefined queries and reports to meet the stated needs of the Aviation Detailers.

A. REQUIREMENTS ANALYSIS

The major requirements data gathering activities consisted of interviews conducted with the Aviation Detailers Branch Head, as well as detailers and allocations specialists within BUPERS. The shortcomings revealed during these interviews have been discussed in Chapter III. From these meetings the following database requirements emerged:

- Depict current reutilization throughout the Navy
- Depict reutilization for each community
- Depict reutilization for each NEC, Paygrade and/or Rating
- Depict reutilization for the personnel under orders by these same breakdowns
- Depict non-reutilization (i.e. where sailors are being detailed in lieu of being reutilized and why)
- Provide these tools in a form that can be used easily from a desktop

B. LOGICAL DATABASE DESIGN

Given these requirements, we decided to construct a relational database in Microsoft Access. To initiate the database construction process, a semantic diagram was created using Salsa™ (Figure 8).

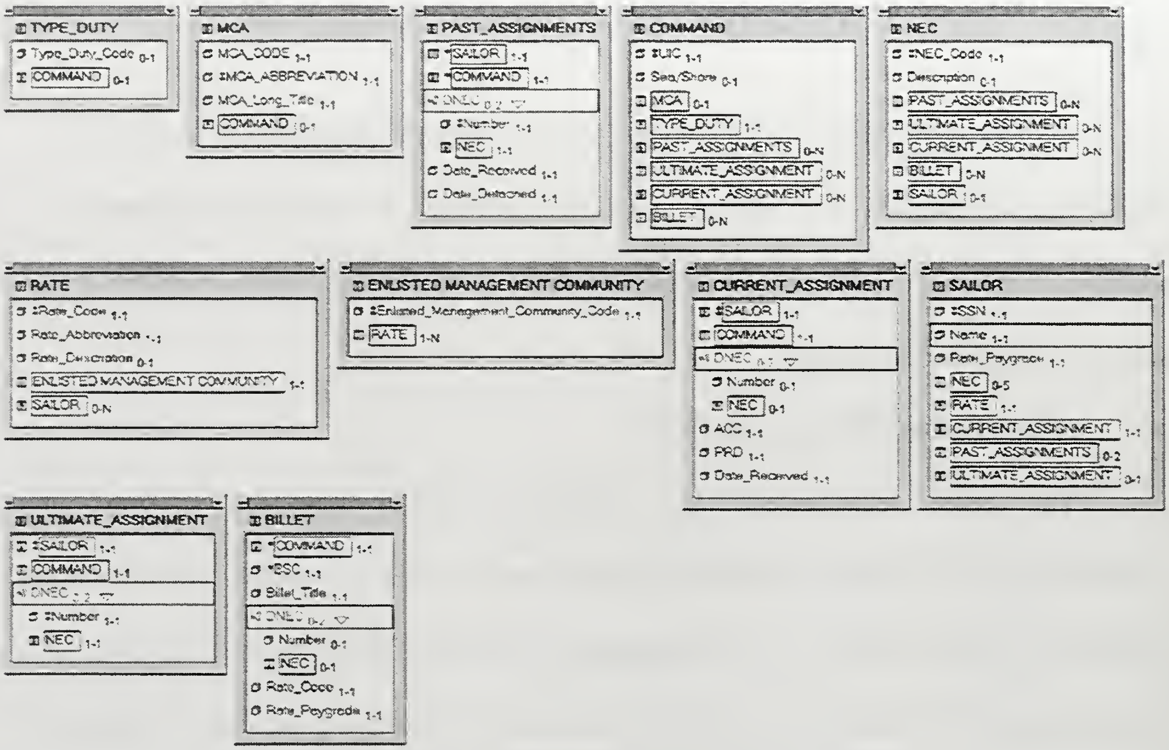


Figure 8. Semantic Object Model

Salsa is based on the semantic object model, a database implementation technique that enables users to design data models according to the way they normally think of information, rather than traditional table/field designs. Each of the Salsa objects is described in more detail below.

The focal object shown in Figure 9 is the Sailor. The unique identifier for each sailor is his or her SSN. Each Sailor in our model has one Name, one Paygrade, one RATE, One CURRENT_ASSIGNMENT, zero to five NECs, zero to two PAST_ASSIGNMENTS, and zero or one ULTIMATE_ASSIGNMENT. The fields that are blocked in and capitalized indicate a separate object attribute that defines a relationship between the host object and another object. DNEC information for each sailor is contained in the CURRENT, ULTIMATE and PAST ASSIGNMENTS.

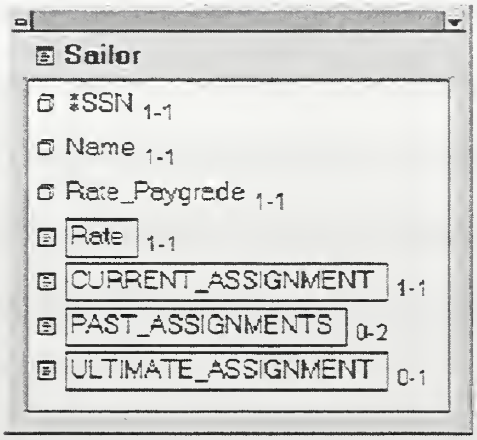


Figure 9. Category SAILOR from Semantic Model

The object ULTIMATE_ASSIGNMENTS represents the data concerning the future command of a sailor with orders. The unique identifier for ULTIMATE_ASSIGNMENTS is a combination of SAILOR and COMMAND. Each Sailor in our model can have zero or one set of orders, so SAILOR is unique in this object. Each ULTIMATE _ASSIGNMENT object has one COMMAND, and zero to two DNECs.

The object PAST_ASSIGNMENTS represents the past history of the sailor. The unique identifier is the combination of SAILOR and COMMAND and DATE. Each PAST_ASSIGNMENT has a report date, detach date, and zero to two DNECs.

The object CURRENT_ASSIGNMENT represents the command to which the sailor is currently attached. The unique identifier is the SAILOR because each sailor has one and only one CURRENT_ASSIGNMENT. Each CURRENT_ASSIGNMENT has a Date_Received, PRD, zero to two DNECs, and an ACC code. The ACC code represents the assignability of a particular sailor.

The object COMMAND represents each command within the Navy. The unique identifier is the UIC. Each COMMAND has one Sea/Shore designation, one TYPE_DUTY, zero to many PAST_ASSIGNMENTS, zero to many CURRENT ASSIGNMENTS, zero to many ULTIMATE _ASSIGNMENTS, and zero to many BILLETS. The assignment objects represent sailors that have been, are, or will be assigned to each command.

The object NEC represents a specific skill set that is defined by the Navy. The unique identifier is the NEC_Code. There is one description for each NEC. There are zero to many PAST_ASSIGNMENTS, CURRENT_ASSIGNMENTS and ULTIMATE_ASSIGNMENTS that are the link to the DNECs that sailors have been assigned. There are zero to many billets that require this DNEC, and there are zero to many SAILORS that possess this NEC in their skill set.

The object RATE represents a SAILORS rating designation within the Navy. This is their field of expertise. The unique identifier is the RATECODE. Each RATE has a description, an abbreviation, belongs to one ENLISTED_MANAGEMENT_COMMUNITY and has zero to many SAILORS that hold that RATE.

The object TYPE_DUTY represents the platform type of a command. Its unique identifier is the Type_Duty_Code. There is one description and there are zero to many commands of each TYPE_DUTY.

The object ENLISTED MANAGEMENT COMMUNITY represents the community that each rate belongs to. The unique identifier is the Enlisted_Mangement_Code. It has one to many RATES that belong to it.

Once the semantic model is completed, the next step is to create a relational database model. This can be generated automatically by Salsa for Microsoft Access, and then modified as needed. This results in a diagram of the relational database including all the tables in the database and the relationships amongst them. Once the tables and relationships have been specified initially, the next step is to normalize the tables to avoid undesirable update anomalies.

1. Normalization

When creating relationships between tables and data, care must be taken to ensure that key fields are unique and anomalies do not occur. Keys sometimes consist of a group of attributes. For example, in the table in Figure nn, sailors can hold more than one NEC, so the same SSN value can occur in more than one row. Therefore a key must be made from both SSN and NEC to allow for sailors to hold multiple NEC's. However, an anomaly can occur when we delete sailor 111-11-1111 from table (a) of Figure 10, We would lose information about the existence of both the NEC 9502 and the title Instructor. These are examples of anomalies that can be avoided if we normalize the relations properly. In the example above we would separate the data into two tables, one for SSN and NEC, and one for NEC and Title (see tables (b) of Figure 10).

(a)

SSN	NEC	NEC Title
111-11-1111	9502	Instructor
111-11-1111	6258	Technician
222-22-2222	6234	Mechanic
333-33-3333	6547	Engineman
444-44-4444	6234	Mechanic

(b)

SSN	NEC
111-11-1111	9502
111-11-1111	6258
222-22-2222	6234
333-33-3333	6547
444-44-4444	6234

NEC	NEC Title
9502	Instructor
6258	Technician
6234	Mechanic
6547	Engineman

Figure 10. Examples of Anomalies Within Tables

2. First Normal Form

Any data that meets the definition for a relation is in First Normal Form. The cells must hold only one value, all entries in columns must be of the same data type, each column must have a unique name, and no two rows of the table may be identical. All of the tables created in this database have defined key fields, which are unique, ensuring uniqueness of rows. Also, columns were developed that are of predefined data types and

lengths, and each column contains information from the same domain. All of the tables meet the First Normal Form definition.

3. Second Normal Form

A table is in Second Normal Form when it meets the requirements of First Normal Form and all non-key attributes are dependent on the entire key. In the example in Figure nn, NEC Title is only dependent on the NEC, not on SSN. In this case, Title is only dependent on part of the key field; as a result, when we delete certain SSNs we may inadvertently lose important Title data. To remedy this situation, we would create a separate table for NEC and Title. Tables that have only one key are automatically in Second Normal Form. The only tables that have more than one key field in the database are the SAILOR_NEC table and the COMMAND_BILLETS table. Both of these tables have been decomposed into second normal form.

4. Third Normal Form (3NF)

Third Normal Form meets the requirements of Second Normal Form and has no transitive dependencies, see Figure 11. In the table above SSN is the key field. Rate is dependent on SSN and RateCode is dependent on Rate. Therefore RateCode has a transitive dependency on Rate. If we were to delete the third row in this table we would lose information about both the Rate and the RateCode. The fact that the RateCode for DT is 8600 would be lost. This is because in this example the RateCode has a transitive dependency on SSN, and therefore is not in 3NF. We must remove this dependency to correct the situation.

Normalization often operates in opposition to database efficiency, thus it is the designer's decision about how extensively to implement normalization techniques. One

place where we chose not to convert to 3NF is the table Command where some of these transitive dependencies have been kept in order to create efficiency in the database. Specifically, there are fields that hold location codes, and a field that holds homeport. The location codes are dependent upon the homeport, and if the homeport changes, then all of these fields would need to be updated individually. As a matter of efficiency, the decision was made to keep all of this information in the same table.

SSN	Rate	RateCode
111-11-1111	HM	8000
222-22-2222	HM	8000
333-33-3333	DT	8600
444-44-4444	HM	8000

Figure 11. Example of Third Normal Form

5. Fourth Normal Form

Fourth Normal Form is defined as being in Third Normal Form and having no multi-valued dependencies. Figure 12 demonstrates Fourth Normal Form. A sailor can have more than one NEC and more than one DNEC. There is no logical way to maintain this in one table. The primary key would be either (SSN, NEC) or (SSN, NEC, DNEC). It would appear that DNEC and NEC would be associated with each other, even when they are not. A sailor can hold two NECs and no DNECs, and vice versa. In order to correct this we separate the NEC and DNEC data into their own tables. The primary keys are (SSN, NEC) and (SSN, DNEC) respectively.

SSN	NEC	DNEC
111-11-1111	9502	6258
111-11-1111	3246	
222-22-2222		9502
222-22-2222		6547

SSN	DNEC
111-11-1111	6258
222-22-2222	9502
222-22-2222	6547

SSN	NEC
111-11-1111	9502
111-11-1111	3246

Figure 12. Fourth Normal Form

C. PHYSICAL DATABASE DESIGN

The DBMS engine we have chosen for the development of the relational table structure for the reutilization data is Microsoft Access 2000. The tables are an extension of the ER diagram in Figure 13.

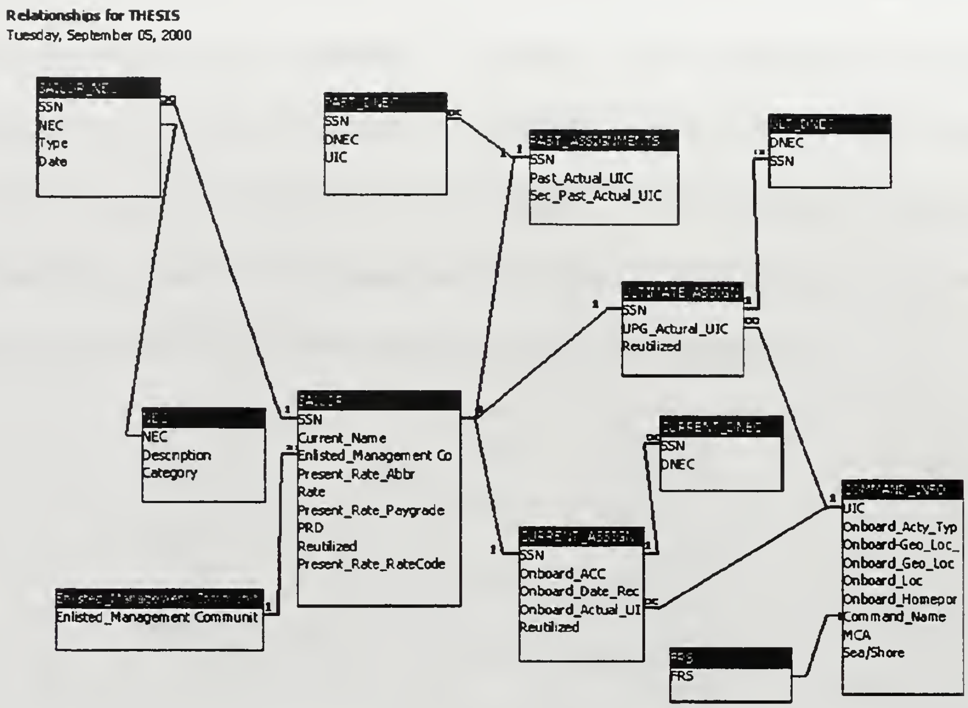


Figure 13. Entity Relationship Diagram (Microsoft Access)

D. DATA SOURCE EXTRACTION PROCESS

To populate the database, a list of fields from the EMF was provided to EPMAC, which was then able to create and transfer a full extract with all of the fields included. This extract was provided in a large Microsoft Access table, EMF200004, the structure of which is shown by table X. This same table can be provided by EPMAC and used to populate the database using the following queries. The prototype table was an extract from April 2000, and holds over 309,000 enlisted records.

Columns

Name	Type	Size
SSN	Text	9
Current_Name	Text	27
Enlisted_Management Community	Text	4
Present_Rate_RateCode	Text	4
Present_Rate_Paygrade	Text	1
Present_Rate_Rate_Abbreviation	Text	5
PNEC	Text	4
SNEC	Text	4
TNEC	Text	4
OTNEC	Text	4
ONNEC	Text	4
OnBoard_Parent_UIC	Text	5
OnBoard_Actual_UIC	Text	5
OnBoard_Acty_Name	Text	16
DNEC1	Text	4
DNEC2	Text	4
Past_Parent_UIC	Text	5
Past_Actual_UIC	Text	5
Past_Acty_Name	Text	16
Past_DNEC1	Text	4
Past_DNEC2	Text	4
Sec_Past_Parent_UIC	Text	5
Sec_Past_Actual_UIC	Text	5
Sec_Past_Acty_Name	Text	16
Second_Past_DNEC1	Text	4
Second_Past_DNEC2	Text	4
UPG_Parent_UIC	Text	5
UPG_Actural_UIC	Text	5
UPG_Acty_Name	Text	16
UPG_Asgn_DNEC1	Text	4
UPG_Asgn_DNEC2	Text	4
Primary_NEC_Date	Text	4
Secondary_NEC_Date	Text	4
Tertiary_NEC_Date	Text	4
Quaternary_NEC_Date	Text	4
Quinary_NEC_Date	Text	4
UPG_Asgn_Rate_Reason_1	Text	1
UPG_Asgn_Rate_Reason_2	Text	1
Onboard_Asgn_Rate_Reason_1	Text	1
Onboard_Asgn_Rate_Reason_2	Text	1
DNRC1	Text	4
DNRC2	Text	4
PRD	Text	6

Table 6. EMF200004

Billet data was also imported into the database to fill the Billets table. This data was contained in a flat file similar to the one above. The source of the billet data was TFFMS.

1. Queries

Queries are created to populate the database, and to create reutilization data after all of the data is imported into the EMF200004. The Sailor table is populated with:

```
SELECT EMF200004.SSN, EMF200004.Current_Name, EMF200004.PRD,
        EMF200004.[Enlisted_Management Community],
        EMF200004.Present_Rate_RateCode, EMF200004.Present_Rate_Paygrade
        INTO SAILOR
 FROM EMF200004;
```

All of the tables are populated with similar queries. The SAILOR_NEC table is populated using five different queries, in order to hold all of the NECs in one table, with the primary key being (SSN,NEC). The query for Primary NEC (PNEC) is:

```
SELECT EMF200004.SSN, EMF200004.PNEC,
        EMF200004.Primary_NEC_Date, "PNEC" AS Type
INTO SAILOR_NEC
FROM EMF200004
WHERE (([EMF200004].[PNEC])<>"0000");
```

The billet information imported from TFMMS is used to populate the COMMAND table, the BILLETS table and the TYPE_DUTY table. They are populated in the same manner as above.

The queries are then created to determine reutilization. The Reutilized fields in CURRENT_ASSIGNMENT and ULTIMATE_ASSIGNMENT are modified to hold a '1' if the sailor is being reutilized. This information can be used to create tables that can be exported to an OLAP tool for multidimensional analysis (see next Chapter). The query used is:

```
UPDATE ((Sailor INNER JOIN CURRENT_DNEC ON Sailor.SSN =
        CURRENT_DNEC.SSN) INNER JOIN PAST_DNEC ON
        (CURRENT_DNEC.DNEC = PAST_DNEC.DNEC) AND
        (CURRENT_DNEC.SSN = PAST_DNEC.SSN)) INNER JOIN Sailor_NEC
        ON (PAST_DNEC.SSN = Sailor_NEC.SSN) AND (PAST_DNEC.DNEC =
        Sailor_NEC.NEC)
SET     Sailor.Reutilized = 1;
```

The same type of query is used to update the reutilized field of ULTIMATE ASSIGNMENT.

The following sections explain the main tables and relationships of the relational model.

2. Sailor Table

The first table created represents each individual Sailor, Table 7. The primary key (PK) of this table is SSN. SSN is a unique identifier for every sailor. Included in this table are name, rate, paygrade, and rate abbreviation. These attributes and those that are related to SSN will be used to determine if the sailor is being reutilized in their current or ultimate billet.

SSN	Current_Name	Enlisted Management Community	Present_Rate_Abbr	Rate	Present_Rate_Paygrade	PRD	Reutilized	Present_Rate_RateCode
xxxxxxxxx	LEWIS LOCHLAIN PUTNAM	E200	BMC	BM	7	200109		0100
xxxxxxxxx	YOUNG KEVIN JAY	B340	STG1	STG	6	200101		0401
xxxxxxxxx	MANK BRIAN ERNEST	B420	ETC	ET	7	200205		1000
xxxxxxxxx	FOOTE KEVIN M	B400	BM2	BM	5	200204		0100
xxxxxxxxx	CRITTENDEN BRIAN PATRICK	B420	ET1	ET	6	200203	1	1000
xxxxxxxxx	BERNIER JILL ELIZABETH	G027	HM2	HM	5	200302	1	8000

Table 7. Extract of SAILOR table

3. Current Assignment Table

Table 8 represents the current assignments of each individual Sailor. It contains all of the information on the Sailor's current assignment, including DNEC, date of receipt and UIC. DNEC information is compared with the NECs a Sailor holds and the DNECs used at previous commands to determine reutilization. It is linked to the Command

Table, which holds information on all of the commands. The unique identifier for this table is SSN, as each Sailor can have only one current assignment.

SSN	Onboard_ACC	DNEC1	DNEC2	Onboard_Date_Rec	Onboard_Actual_UIC
xxxxxxxxx	100	5336		980819	43504
xxxxxxxxx	100			980129	45196
xxxxxxxxx	100	0000		990416	60087

Table 8. Extract of CURRENT_ASSIGNMENT table

4. Past Assignments Table

Table 9 represents the Sailor's past assignments. A Sailor can have zero, one or two past assignments, but the EMF only maintains data for the previous two. The table contains DNEC information and date information. Since the table only holds up to two previous commands, the date information may be used to determine how long the history is.

SSN	Past_Actual_UIC	Past_DNEC1	Past_DNEC2	Sec_Past_Actual_UIC	Second_Past_DNEC1	Second_Past_DNEC2
xxxxxxxxx	30215	5337		82631	0000	
xxxxxxxxx	62443	9585		62443		
xxxxxxxxx	21531	1425	1452	30811		
xxxxxxxxx	20014	0000		21109	0000	

Table 9. Extract of PAST_ASSIGNMENTS table

5. Ultimate Assignment Table

Table 10 represents the Sailors that have received orders to their next command. It contains DNEC information, which can be used to determine reutilization for Sailors who are currently under orders.

SSN	UPG_Actural_UIC	UPG_Asgn_DNEC1	UPG_Asgn_DNEC2	Reutilized
xxxxxxxxx	08321	8404		
xxxxxxxxx	32770	0421		1
xxxxxxxxx	61564	0000		
xxxxxxxxx	66045	8342	9502	1

Table 10. Extract of ULTIMATE_ASSIGNMENT table

6. Sailor NEC Table

Table 11 contains all the NECs that each Sailor holds. A Sailor can hold up to five NECs. The unique identifier for this table is the composite key made up of SSN and NEC. It holds the date the NEC was awarded and the specific precedence of it (Primary, Secondary, Tertiary . . .)

SSN	NEC	Type	Date
xxxxxxxxx	5337	SNEC	9511
xxxxxxxxx	9502	TNEC	9301
xxxxxxxxx	0505	PNEC	9709
xxxxxxxxx	9585	SNEC	9411
xxxxxxxxx	1428	PNEC	9603

Table 11. Extract of SAILOR_NEC table

7. Onboard and Under_Orders Tables

Two tables are then created to export to the OLAP tool, ONBOARD and UNDER_ORDERS. They represent reutilization data for the Navy as a whole, and those individuals under orders. The following drill-downs of information were specifically requested from the Aviation detailers:

- Rating-Paygrade-Nec-SSN-Previous Tour/Experience
- ECM-Rating-NEC-Paygrade-SSN

- PlatformType-Rating-NEC-paygrade-SSN
- Platform Type-MCA-Sea/Shore-UIC-Rating-NEC-Paygrade-SSN
- Platform Type-MCA-UIC-Rate-NEC-Paygrade-SSN

All of this information was compiled into the two tables to facilitate the creation of multidimensional data cubes for use in reporting reutilization information. These tables are the result of queries created in ACCESS. Either the tables or the queries could be used in the OLAP tool, either would represent the same data. PRD information was included in the ONBOARD table in order to list the personnel that would be available for orders within a specific timeframe. The PRD field is not important once a person has received orders. DNEC information was included in the UNDER_ORDERS table for the upcoming command. This was used to show non-reutilization. For example, it could show where aviation personnel are going if they are not going to Aviation billets.

Table 12 contains the fields of the table ONBOARD exported to the OLAP tool. Table 13 is the field of the table UNDER_ORDERS exported to the OLAP tool.

Table: ONBOARD Page: 1

Columns

Name	Type	Size
SSN	Text	9
PRD	Long Integer	4
Enlisted_Management Community	Text	4
Rate	Text	4
Present_Rate_Paygrade	Integer	2
MCA	Text	19
Sea/Shore	Text	28
Onboard_Actual_UIC	Text	5
Reutilized	Integer	2
Onboard_Acty_Type	Text	4
NEC	Text	4
Type	Text	5
NEC/DNEC	Text	4
DNEC	Text	4

Table 12. Table ONBOARD

Table: UNDER_ORDERS Page: 1

Columns

Name	Type	Size
SSN	Text	9
Enlisted_Management Community	Text	4
Rate	Text	4
Present_Rate_Paygrade	Integer	2
MCA	Text	19
Sea/Shore	Text	28
UPG_Actural_UIC	Text	5
UPG_Asgn_DNEC1	Text	4
UPG_Asgn_DNEC2	Text	4
Reutilized	Integer	2
Onboard_Acty_Type	Text	4
NEC	Text	4
Type	Text	10
Aviation_DNEC	Text	3
NEC/DNEC	Text	4

Table 13. Table UNDER_ORDERS

E. REPORTS

Reports were created in ACCESS to demonstrate the abilities of the relational database. These reports need to be further refined to meet all of the needs of the manager using the information. Once these reports are created, a manager can print them out for every dataset that is used to populate the database. Further, output can be limited to

individual manager selections, such as a list of personnel within a certain PRD, or a list of personnel holding a specific rate and NEC. There are many reports that can be created using this tool, which would be extremely useful for the specific purpose of reutilizing NEC's.

A report was created to show the reutilization of all personnel that hold the 83XX NEC and have used it within the current or previous two commands. The personnel totals on this report include only personnel that are under orders and possess the experience in these NECs. In this particular example, it is assumed that if a sailor has experience in any 83XX and has orders to an 83XX billet, that sailor is being reutilized. Figure 14 shows a page of the 83XX Orders Report.

Currently, managers are using a spreadsheet created manually to list personnel under orders to Fleet Replacement Squadrons (FRS). A report was created in ACCESS to detail the same information in a similar format. This report is shown in Figure 15.

83xx_ORDERS

RateCode	Activity Type	Experience	Reutilized	
AD				
	GDTY	116	24	21%
	HC	6	4	67%
	HS	2	2	100%
	HSL	8	7	88%
	MFA	3	2	67%
	NOC	6	2	33%
	STF	3	1	33%
	VAQ	4	2	50%
	VF	1	1	100%
	VFA	3	3	100%
	VP	8	4	50%
	VRC	4	4	100%
	VX	6	5	83%
Summary for 'Present_Rate_RateCode' = 6200 (13 detail records)				
		170	*61*	36%
AT				
	GDTY	73	24	33%
	HC	3	3	100%
	HM	1	1	100%
	HS	1	1	100%
	HSL	3	2	67%
	MFA	1	1	100%
	NOC	12	3	25%
	STF	5	2	40%
	VF	8	6	75%
	VFA	1	1	100%
	VP	4	3	75%
	VRC	1	1	100%
	VS	1	1	100%
	VX	1	1	100%
Summary for 'Present_Rate_RateCode' = 6300 (14 detail records)				
		115	*50*	43%
AO				
	GDTY	46	12	26%
	HSL	5	3	60%
	NOC	2	1	50%
	STF	15	9	60%
	VAQ	1	1	100%
	VF	4	2	50%
	VP	3	2	67%
	VX	1	1	100%
Summary for 'Present_Rate_RateCode' = 6500 (8 detail records)				
		77	*31*	40%
AE				
	GDTY	60	20	33%
	HC	4	3	75%
	HM	1	1	100%

Monday, September 04, 2000

Figure 14. 83XX ORDERS

FRS ORDERS

SSN	*RATE*	*DNEC*	*NEC*	*A UIC*	*CMD*	*PAST1*	*PAST2*	*ULT_CMD*
xxxxxxxxx	AMSC	8378	7232 8305 8319 8378	03364	CV 64	S NLTU	NAS	HC 3
xxxxxxxxx	AD2	8215	8241 8215 8216	30808	S FLTTC	HC 11	HC 3	HC 3
xxxxxxxxx	AT1	8378	9502 8376	31778	HSL 51 DET	AMTGD	HSL 41	HC 3
xxxxxxxxx	AEC	8378	7136 8377	09192	VS 38	S NLTU	CV 63	HC 3
xxxxxxxxx	AT2	8378	8378 8377	09950	HS 6	HS 10	HS 2	HC 3
xxxxxxxxx	AMH1	8378	8378 8343	55150	HSL 47	S C/C	HC 3	HC 3
xxxxxxxxx	AMH1	8378	8378 8375	42300	HC 11 SEA	HSL 41	HS 2	HC 3
xxxxxxxxx	AMH2	8205	8377 8211 8215	30811	S FLTTC	HC 2	NAS	HC 3
xxxxxxxxx	AMHC	0000	8346 9502 8378 8312	09192	VS 38	AMTGD	VS 41	HC 3
xxxxxxxxx	AMS1	8205	8216 8215	30808	S FLTTC	HC 11	HC 11	HC 3
xxxxxxxxx	AMS2	8878	9585 8378	09951	HS 8	TPU SD	S C/C	HC 3
xxxxxxxxx	AD2	8878	8806	39476	S NAMTD	VRC 30	S CRU	HC 3
xxxxxxxxx	AD1	8378	8378 8345 8319	09950	HS 6	S	COMAE	HC 3
xxxxxxxxx	AE2	8378	8378 8346 8312	09192	VS 38	S	NAS	HC 3
xxxxxxxxx	AE2	8378	8378	55150	HSL 47	HS 10	S C/C	HC 3
xxxxxxxxx	AE2	8378	8378	09372	HS 2	HSL 51	HSL 40	HS 10
xxxxxxxxx	AW2	7876	7876 7815 0812	09164	HS 4	S C/C	S	HS 10

Monday, September 04, 2000

Figure 15. FRS Orders

F. CONCLUSION

A relational database that is created in Microsoft Access provides a relatively flexible tool for reporting information. Many reports can be created and tailored to meet the individual needs of managers using the information. However, it is difficult to manipulate the data dynamically in different ways when addressing specific calculations. For instance, reports can be created to show reutilization among Aviation Rates, or even a specific rate, and a user can select which rates and how many rates for which to show a reutilization rate. However, the user is dependent upon the original static format of the report and has little room for changing the way data is presented. One solution to this limitation could be to create many reports within the relational database, each of which shows the data from a specific perspective.

Another, more dynamic solution that we will explore is the use of an OLAP tool that gives us the expanded ability to depict reutilization from multiple views and subject to multiple manipulations. This expanded capability gives managers a quick method for reporting calculations, specifically with reutilization in mind.

THIS PAGE INTENTIONALLY LEFT BLANK

V. ONLINE ANALYTICAL PROCESSING PROTOTYPE

A. INTRODUCTION

The purpose of the OLAP prototype is to highlight the benefits of using multi-dimensional OLAP analysis to view data accessed from a relational data mart. However, the underlying purpose goes beyond the desire to provide enlisted personnel planners with a tool that will help them perform their jobs. As discussed at the beginning of Chapter II, the enlisted personnel system is used to make decisions that have enterprise wide significance and consequences. We used the term "Critical Success Factor" because the enlisted personnel system must be able to function effectively in a highly dynamic environment; the results of which play a leading role in the successful accomplishment of worldwide naval missions. Although the current enlisted personnel database system often gets the job done, we believe a relational data warehouse or data mart that is augmented by an OLAP tool can provide enlisted personnel planners with significantly increased analytical capabilities. This can, in turn, lead to better decision-making and an increased ability to focus on critical business drivers.

We chose Cognos PowerPlay™ as our OLAP application for this prototype primarily because of its availability. Cognos is one of the market share leaders in the OLAP software industry. Cognos PowerPlay™ is a DOLAP application that can access source data from a relational or non-relational database. PowerPlay™ works by creating multidimensional PowerCubes that contain the dimensions and measures needed to perform data analysis. PowerPlay™ can be deployed on either a standalone PC or as a fully networked application that can operate in a Web-based environment. Figure 16

provides a view of a comprehensive enterprise level deployment. Some of the elements in this architecture are described below:

- **Decision Stream:** Build and deploy a series of linked, subject-area data marts.
- **Transformation Server:** Runs on Windows 98, NT or UNIX to model and build PowerCubes.
- **Architect:** Addresses the need for common enterprise-wide metadata management.
- **Upfront:** A Web portal service that provides end users with a single, Web-based point of personalized access to business intelligence information.
- **Impromptu:** Allows report authors to create complete reporting applications using the capabilities of a Windows client. They then deploy these reports over the Web to any number of users inside or outside of the organization.
- **PowerPlay™ Web:** Enables users to access all leading OLAP servers, for intranet and extranet multidimensional analysis, reporting and report distribution.

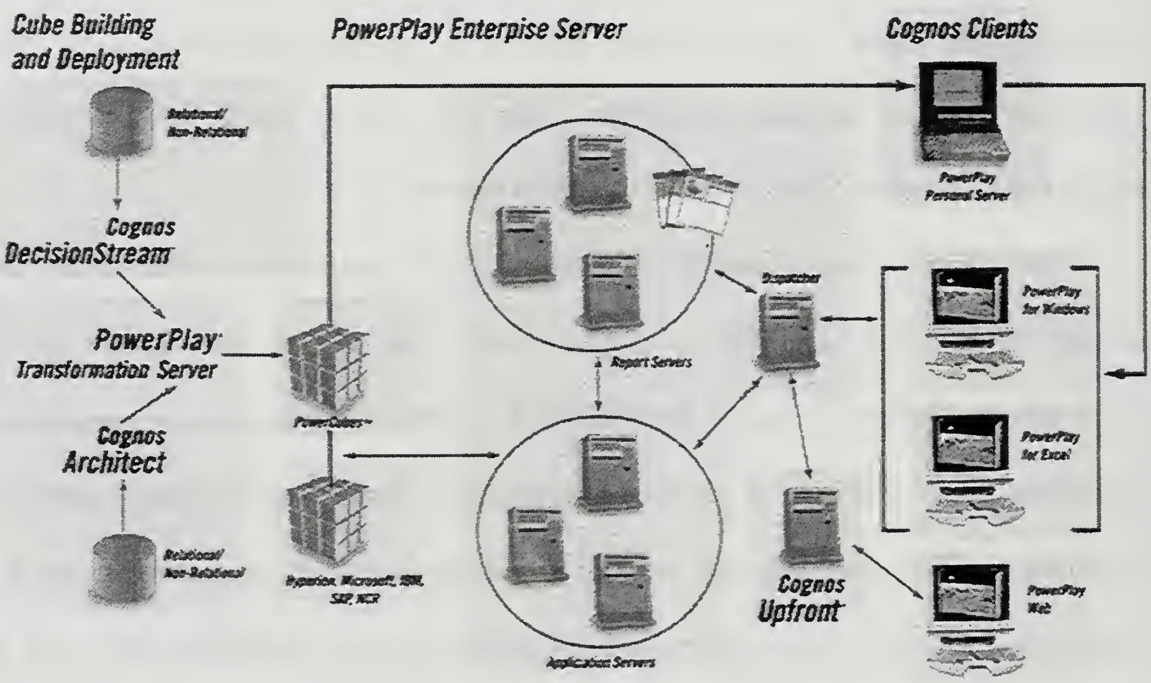


Figure 16. PowerPlay™ Enterprise Server Architecture (Cognos Corp.)

Our goal is to demonstrate the visualization and reporting capabilities of OLAP and how these capabilities can help enlisted personnel managers perform their duties more effectively and efficiently. The OLAP prototype will focus on analyzing NEC reutilization using a standalone PC, Transformation Server and PowerPlay™ for Windows.

B. REQUIREMENTS

The use of an OLAP tool requires thorough research of user requirements and business drivers for a successful implementation. In addition, the implementation process necessitates that a number of critical steps be completed which encompass a range of technical, managerial and creative skills.

1. Performance Metrics

Having critical information about business drivers gives knowledge workers the tools needed to make effective decisions. These critical business drivers need to be specified and defined in the form of performance metrics. A *performance metric* defines key performance indicators that are used to evaluate the organization and provide business intelligence.

It must also be decided how the data will be organized to support the analysis needed. For example, a question to be answered is "What drill-down hierarchy will provide the most benefit to the users?" Properly defining each metric and its associated drill-down hierarchy will ensure it is tailored to address and solve a specific business problem. To support performance metric design, a performance metric should be defined using the template in Table 14.

Term	Definition
Definition	The metric definition is a text description fully describing the semantics of a metric and the organizational goals that it supports.
Computational-Procedure	Decision metrics may be calculated as simple arithmetic quantities (e.g., financial ratios), database queries, qualitative assessments, or as outputs from Our goal for this section is to demonstrate some of the ways data can be manipulated with PowerPlay™. more formal models.
Dimension/Units	The dimension and unit of the metric, e.g., Cost/Unit Cost in $.
Thresholds	Each metric has associated performance threshold levels. These thresholds may be manually prescribed, or statistically derived. The thresholds are translated into simple visual cues.
Periodicity	Each metric has a baseline periodicity for which it is measured. It may be every second, minute, hour, day, week, month, or year depending upon what is being measured. The periodicity is the minimum time granularity for which a metric may be represented.
Scale Level	A metric is assumed to have a numerical value of some kind that may be ordinal, cardinal, or interval-based.
Drill-down-Dimensions	Each metric has affiliated attributes that comprise an n-dimensional space for calculating the metric at various levels of aggregation.

Table 14. Performance Metric Template

For the OLAP prototype, we designed two performance metrics (See Tables 15 and 16). The Branch Chief of Aviation Enlisted Assignments (PERS-404) was interviewed to determine the dimensions that would be most beneficial for enlisted personnel assignments in the aviation community.

Term	Performance Metric
Definition	***%NECReutilizationOrders*** – NEC Reutilization is defined as personnel who have satisfied an NEC requirement within the last two commands. This performance metric is used to determine NEC reutilization rates for E-5 to E-7 enlisted personnel in Aviation rates who are under orders to rotate to a new command. This metric will also allow a planner to determine NEC reutilization rates for personnel under orders to both aviation and non-aviation billets. Command level data will report aggregate NEC reutilization for personnel under orders to the command, not actual command data.
Computational-Procedure	Number of personnel with NEC reutilized / Number of billets filled with personnel DNEC'd into the billet
Dimension/Units	Percentage
Thresholds	White background: 50% - 100% Reutilization Spotted background: 32% - 49% Reutilization Black background: < 32% Reutilization
Periodicity	Monthly
Scale Level	Cardinal-based.
Drill-down-Dimensions	Rate, Paygrade, DNEC1, SSN, NEC, ECM, Platform-Types, MCA, Sea/Shore, UIC, Order Type

Table 15. *%NECReutilizationOrders* Performance Metric

Term	Performance Metric
Definition	***%NECReutilizationOnboard*** – NEC Reutilization is defined as personnel who have satisfied an NEC requirement within the last two commands. This performance metric provides a view of NEC reutilization across the aviation-enlisted community for paygrades E-5 through E-7. This metric will allow a user to distinguish between personnel under orders and not under orders. For those not under orders, it will allow a user to conduct an analysis based on the Personal Rotation Date (PRD).
Computational-Procedure	Number of personnel with NEC reutilized / Number of billets filled with personnel DNEC'd into the billet
Dimension/Units	Percentage
Thresholds	White background: 50% - 100% Reutilization Spotted background: 32% - 49% Reutilization Black background: < 32% Reutilization
Periodicity	Monthly
Scale Level	Cardinal-based.
Drill-down-Dimensions	Rate, Paygrade, DNEC1, SSN, NEC, ECM, Platform-Types, MCA, Sea/Shore, UIC, PRD, Orders

Table 16. *%NECReutilizationOnboard* Performance Metric

2. Star Schemas, Fact Tables, Dimensions, and Hierarchies

Prior to discussing the prototype specifics, it is important to clarify some terms concerning Star Schemas and Cognos PowerPlay™. A *Star Schema* is the primary design mechanism for designing OLAP data structures as discussed in section B-3 in Chapter II. A *Fact Table* is the core data element being analyzed. *Dimensions* are associated with a Star Schema and are defined as attributes about a Fact Table. For example, a retail business data model could have the dimensions Product, Location and Time. Each of these dimensions has attributes. The Time dimension may have the attributes Year,

Month and Week. The fact table in this example could be called SALES. The SALES Fact Table relates the dimensions to the measure of interest, i.e., Sales. (Ramakrishnan, Gehrke, 2000)

Cognos also uses the term Dimension to describe part of the Transformer™ OLAP Model (See Section D-2). In Transformer™, a dimension is a defined as a broad grouping of data that represents major segments of business information. These Dimensions consist of a *hierarchy* of levels or vertical drill-down paths that contain categories. The categories are the operational details of an organization such as command types or paygrades. The drill-down path is the order in which you allow users to drill-down through the various dimensions. For example, a Paygrade Dimension may define the subgroups of E1-E4, E5-E6, and E7-E9. When a user is drilling-down by Paygrade, they can go down a further level into one of the 3 subgroups. Each dimension generally has one to five drill-down levels. They often lead from the highest-level categories in the hierarchy to the lowest, i.e., Year to Month to Week.

3. Time Dimension

A dimension that is always used in a data warehouse and OLAP model is the *Time Dimension*. This is true even if a Time dimension is not explicitly included. The *%NECReutilizationOrders* performance metric does not include an explicit Time dimension because it is designed to take a snapshot of the current operational data. In other words, a Time dimension is irrelevant since we are using monthly data without further time breakdowns. This is often called a *point-in-time* analysis. However, more often than not, data needs to be analyzed at the yearly, monthly, weekly or quarterly level. The *%NECReutilizationOnboard* performance metric uses a time dimension for

analyzing data based on PRD's. To accomplish this, the source data mart includes a PRD date field and the OLAP model includes a Time Dimension. OLAP models are discussed in section D-2 of this chapter.

4. Data Sources

One of the greatest strengths of modern OLAP tools is their ability to use data stored in different formats. For example, Cognos Transformer™ can handle a variety of data formats including Microsoft Access tables or queries, dBase tables, Paradox tables, FoxPro tables, Lotus 1-2-3 database, Excel, and delimited or fixed-field ASCII text. In addition, Transformer™ data for an OLAP model can come from more than one source. This flexibility is built into most OLAP tools and dramatically decreases the likelihood of interoperability problems traditionally associated with using data of varying formats in enterprise level systems.

Another important consideration is how to handle source data changes. Source data often changes regularly, for example, operational data might change minute-by-minute or monthly. Most OLAP tools have scheduler programs to refresh the multidimensional cubes automatically or provide easy commands for manually updating the cubes.

C. STAR SCHEMA DESIGN

Figures 17 and 18 provide a representation of the star schemas designed for the OLAP prototype. In each case, the fact table (hub) is joined to the dimension tables (spokes). Each dimension has a set of associated attributes that pertain to the dimension. For example, the *Sea/Shore* dimension has the attributes *U.S. Sea-based duty* and

Overseas Sea-based duty as well as others. Each of the star schemas origin can be traced back to the performance metric of the same name.

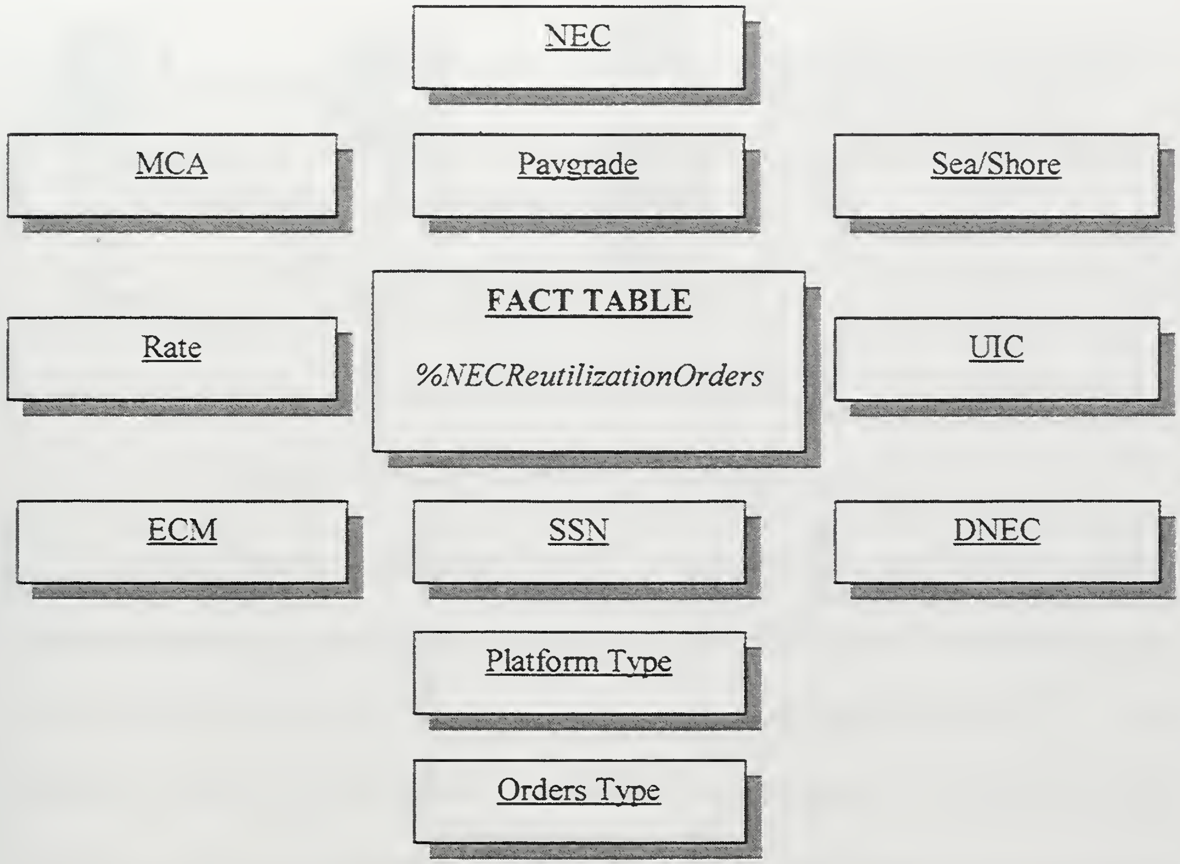


Figure 17. *%NECReutilizationOrders* Star Schema

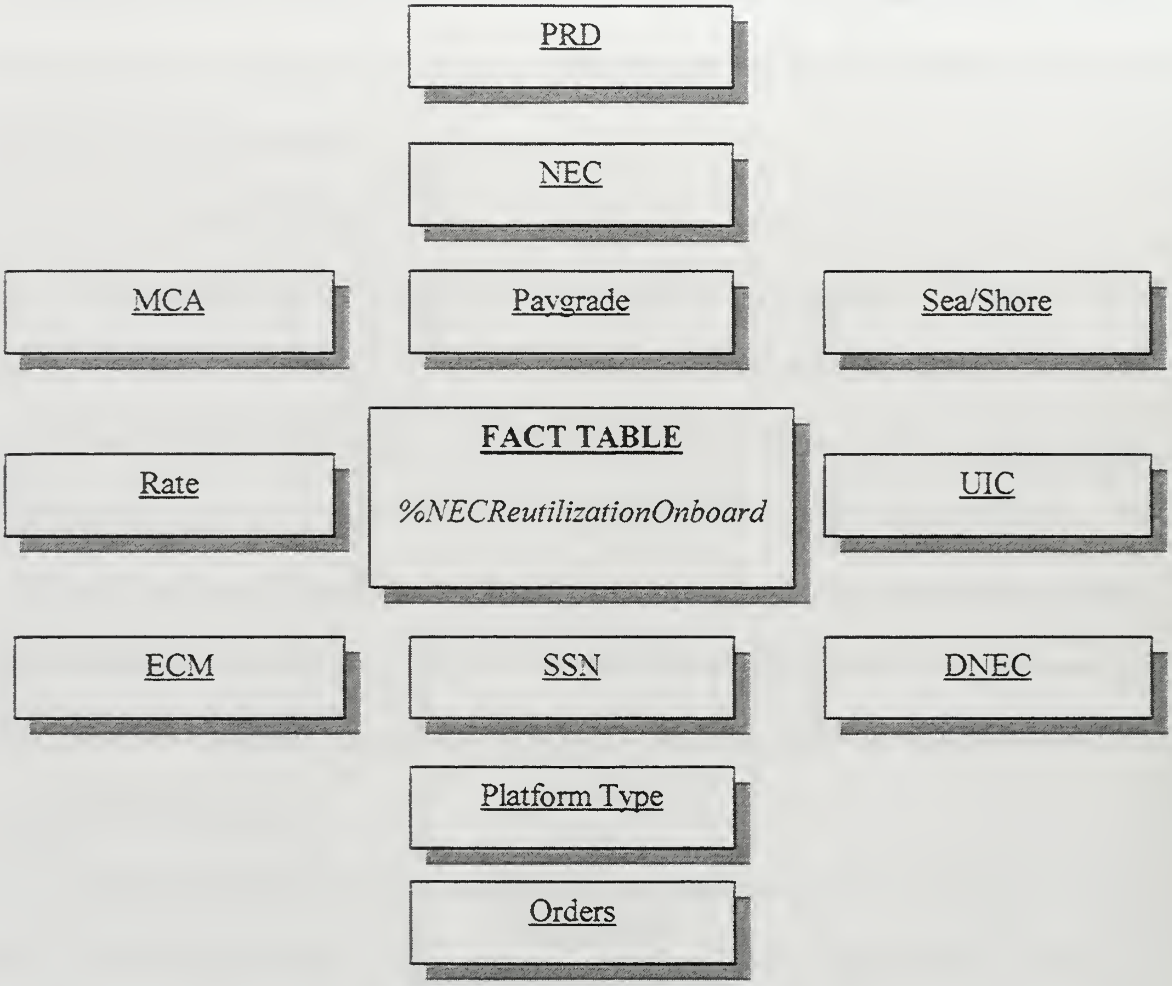


Figure 18. *%NECReutilizationOnboard* Star Schema

D. IMPLEMENTATION

1. Cognos Architecture

The PowerPlay™ architecture used in our prototype consists of the following applications:

- **Cognos Transformer™** – Structures data from various sources into a model used to create multi-dimensional PowerCubes.
- **Cognos PowerPlay™ for Windows**- Graphical user interface that provides access to the PowerCube and the ability to navigate through the data to discover patterns, trends and information to make decisions.

Figure 19 displays the PowerPlay™ architecture used in our OLAP prototype.

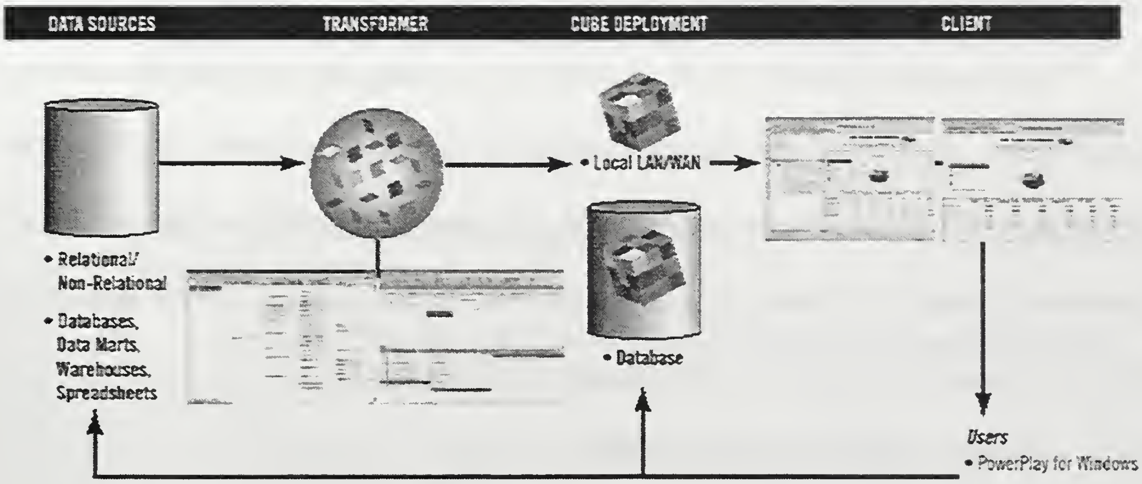


Figure 19. Cognos PowerPlay™ Architecture (Cognos Corp.)

The **data source** is the Microsoft Access relational data mart of Navy enlisted personnel data that was developed in Chapter IV. Transformer™ will structure the data into OLAP models that are used to create multidimensional PowerCubes. **Cube deployment** occurs on the local PC. The **client** is PowerPlay™ for Windows, which provides the interface between a user and a PowerCube, and enables a user to perform multi-dimensional data analysis.

2. Cognos Transformer™ OLAP Modeling

An OLAP model is used to develop the structure of a multidimensional cube. It contains the dimensions, hierarchies and measures that are needed to solve a business problem. Performance metrics are the architectural blueprints used to build an OLAP model from the source data. For the NEC Reutilization OLAP prototype, the process starts by importing a Microsoft Access database table from the prototype relational data mart developed in Chapter IV into Transformer™. The Transformer™ OLAP Model allows the creation and deletion of dimensions, hierarchies and measures in order to meet

the requirements of a performance metric. The only truly limiting factor encountered in OLAP model design is the data itself.

The dimensions roughly follow the common technique of structuring the vertical hierarchy with data from the highest to lowest level. A benefit of using the Transformer™ OLAP model or equivalent is that the model can easily be changed if user requirements change. For example, a new dimension can be developed or the drill-down order easily changed to reflect new requirements.

a. *%NECReutilizationOrders*

Figure 20 displays the *%NECReutilizationOrders* OLAP model.

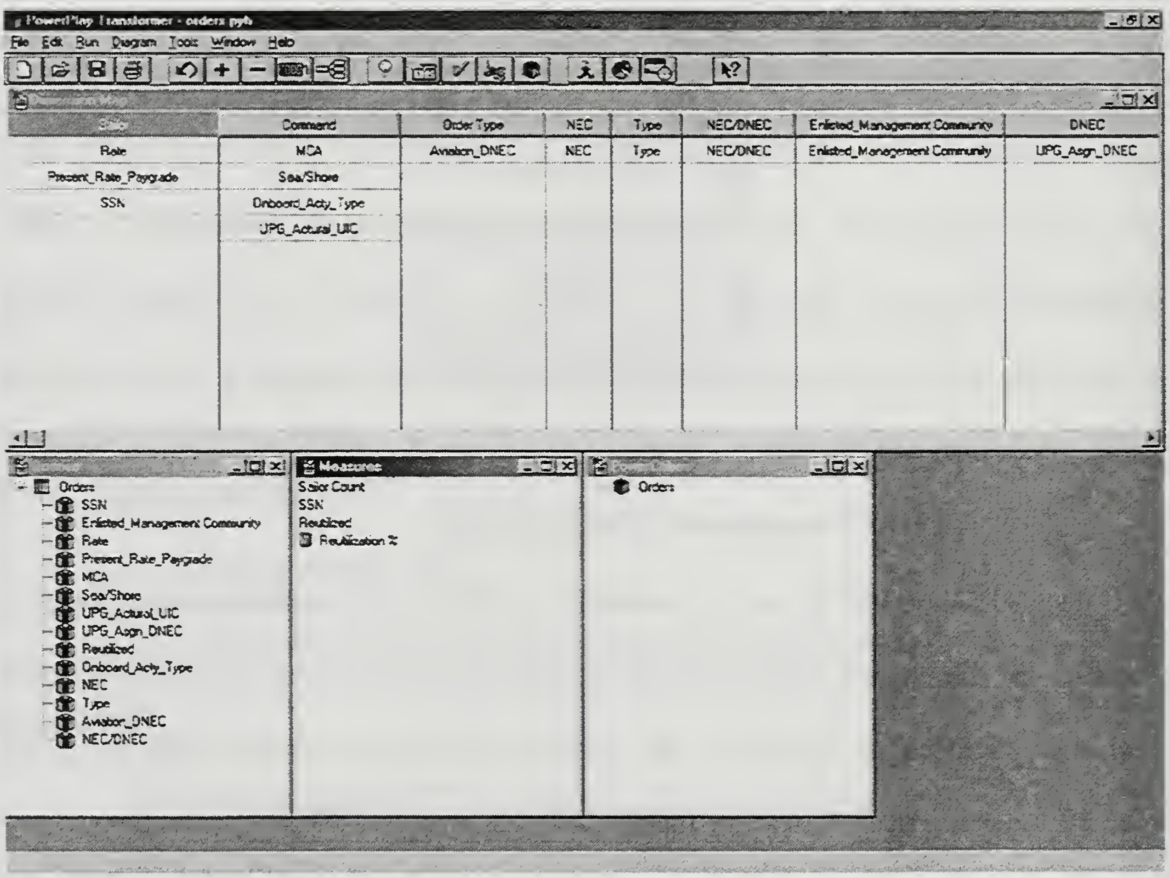


Figure 20. *%NECReutilizationOrders* Transformer™ OLAP Model

Figure 20 includes the following sections:

- **Dimension Map** – This is located in the upper section of Figure xx and contains broad groupings of data that represent a navigation path through one or more dimensions.

- **Queries** – The fields (columns) from the Access table.

- **Measures** – Numerical data used to gauge performance. For this model, the Measures include:

  - Reutilization % – This is a calculated measure that uses the formula: *percent (Reutilized, Sailor Count)*

  - Reutilized – Number of NEC's reutilized based on definition in the performance metric

  - SSN – Total number of personnel DNEC'd

  - Sailor Count – A category count of SSN's

- **PowerCube** – Multidimensional cube formed from the OLAP model.

b. %NECReutilizationOnboard

Figure 21 displays the *%NECReutilizationOnboard* OLAP model. Note that this model has a time dimension called PRD. This will enable planners to analyze data based on PRD's.

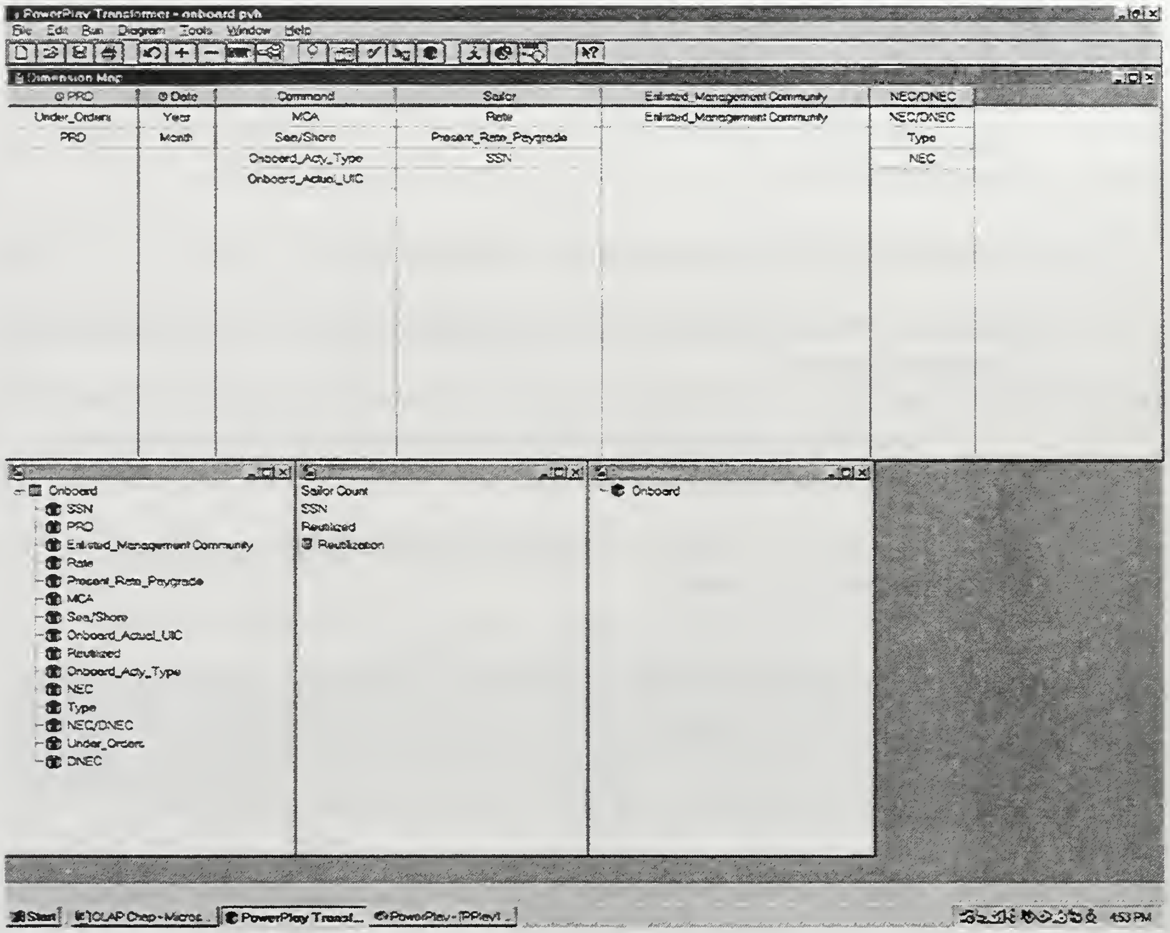


Figure 21. *%NECReutilizationOnboard* Transformer™ OLAP Model

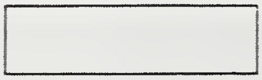
Figure 21 includes similar sections as described for the *%NECReutilizationOrders model.*

E. DATA ANALYSIS – NEC REUTILIZATION

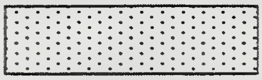
PowerPlay™ provides many different graphical formats for viewing data. For example, pie charts, bar charts, line charts, scatter plots and cross-tabular tables. The proper choice should be based upon user requirements. Color or pattern coding data threshold values are useful ways to view data in an OLAP application. For the prototype we have used distinctive patterns to allow easy identification of NEC reutilization percentage rates. NEC reutilization percentage rates are given three threshold levels:

Excellent, Good and Bad along with corresponding reutilization rates. These are arbitrary and being used for demonstration purposes only. The threshold levels have been designated with the following patterns:

- Excellent – 50 % to 100 % NEC reutilization: White background



- Good – 33% to 49.99% NEC reutilization: Spotted background



- Bad – 0% to 32.99% NEC reutilization: Black background

1. %NECReutilizationOrders Analysis

Our goal for this Subsection is to demonstrate some of the ways data can be manipulated with PowerPlay™. Figures 22 through 27 provide a sampling of PowerPlay™ screen shots of the *%NECReutilizationOrders* performance metric. A benefit of OLAP is that it allows you to easily create a variety of views of the data. Dimensions in the OLAP model that have vertical drill-down paths provide pre-defined flows to view the data. Any dimension can be added by simply dragging and dropping the desired dimensions to build the view within PowerPlay™.

Aviation personnel can have orders to both aviation and non-aviation related billets. Therefore, we chose to give a user the option of analyzing NEC reutilization from both perspectives. Figure 22 displays NEC reutilization percentage rates for aviation personnel under orders to both billet types. As expected, the NEC reutilization rate is much higher for aviation personnel under orders to an aviation billet. Note how the pattern coding applies to both the bar chart and the cross-tabulation display. It also provides a total to display overall NEC reutilization rates. Drilling deeper into the data can be accomplished by simply double-clicking on the bar chart or the cross-tab display.

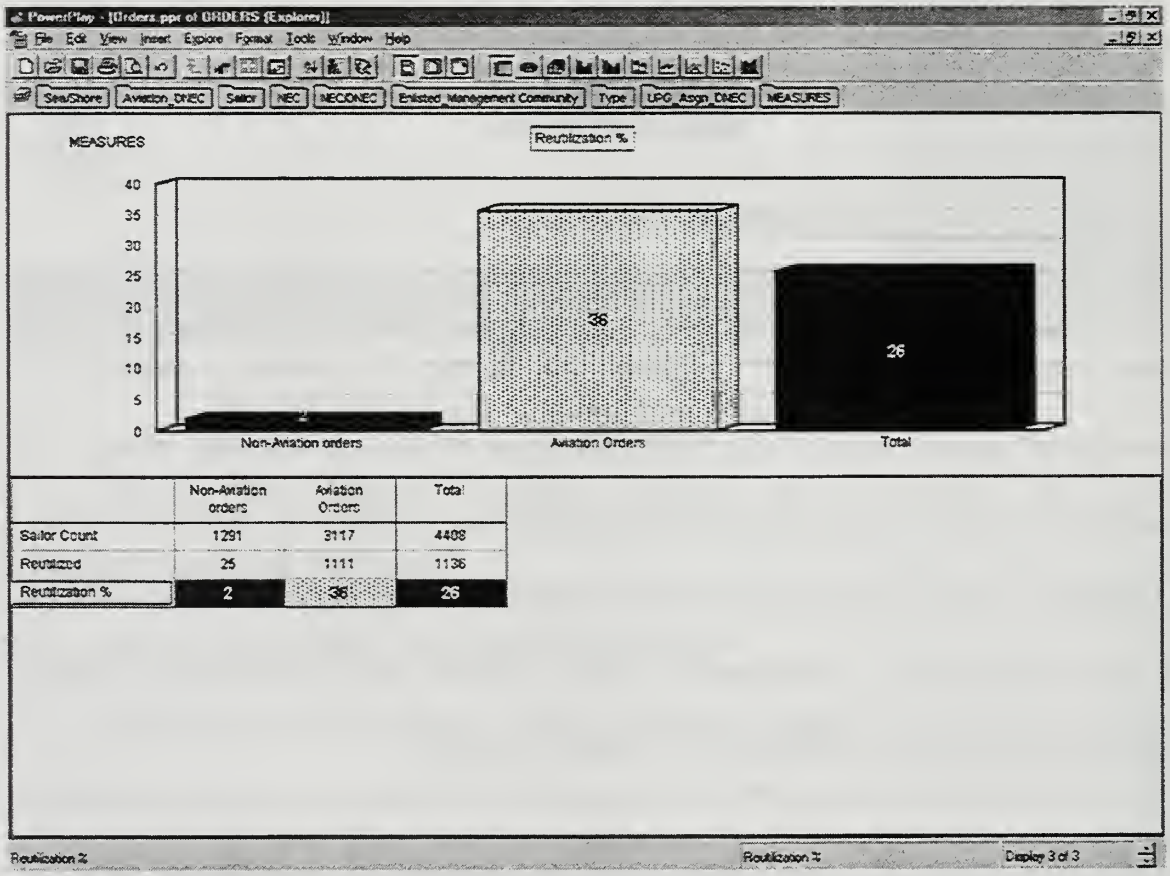
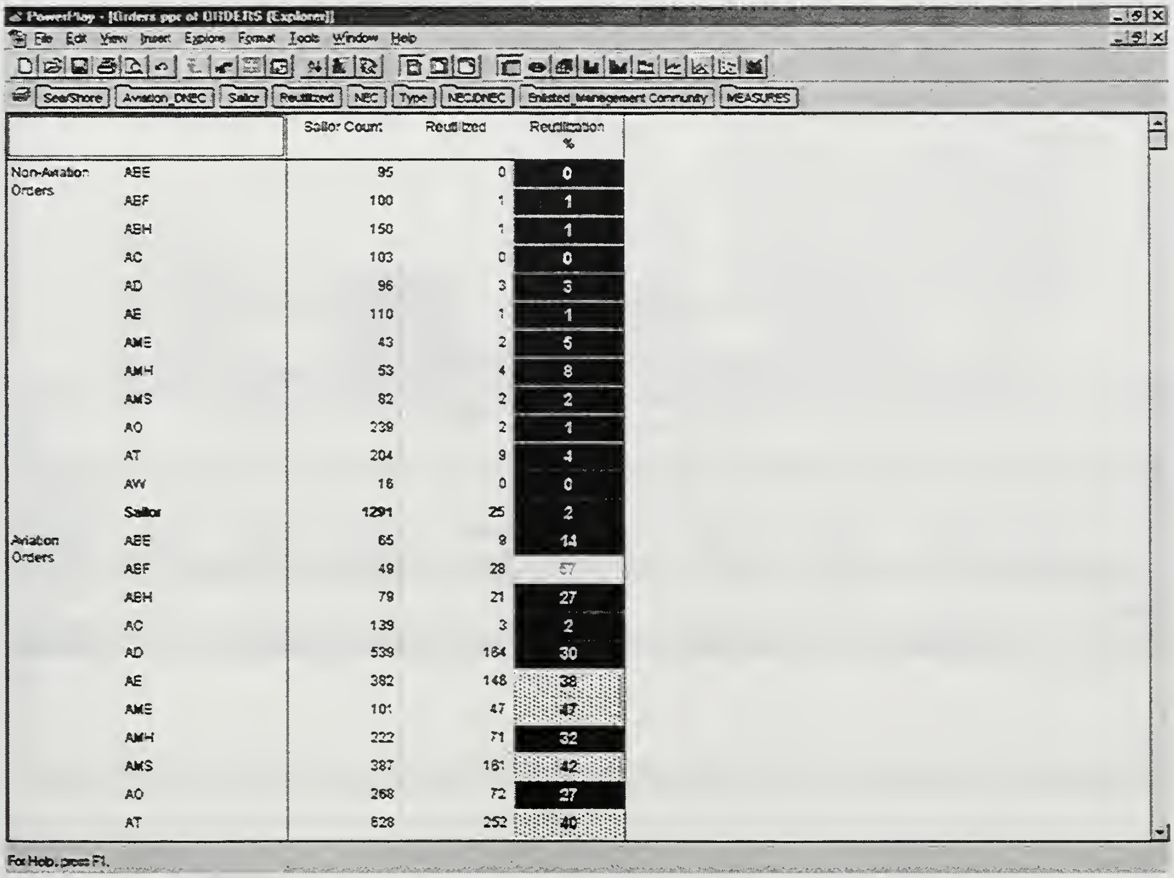


Figure 22. Order Type View of *%NECReutilizationOrders* Performance Metric

Figure 23 is obtained by dragging and dropping the desired dimensions into a cross tabulation display. To obtain this view, the Order Type dimension and Sailor dimension were used. Note again how the pattern coding of the cross-tabulation table provides easy identification of Rates with Bad (black), Good (spotted) and Excellent (white) NEC reutilization rates.



		Sailor Count	Reutilized	Reutilization %
Non-Aviation Orders	ABE	95	0	0
	ABF	100	1	1
	ABH	150	1	1
	AC	103	0	0
	AD	96	3	3
	AE	110	1	1
	AME	43	2	5
	AMH	53	4	8
	AMS	82	2	2
	AO	239	2	1
	AT	204	9	4
	AW	16	0	0
	Sailor	**1291**	**25**	2
Aviation Orders	ABE	65	9	14
	ABF	49	28	57
	ABH	79	21	27
	AC	139	3	2
	AD	539	164	30
	AE	382	148	38
	AME	101	47	47
	AMH	222	71	32
	AMS	387	161	42
	AO	268	72	27
	AT	628	252	40

Figure 23. Rate View of Enlisted Community View *%NECReutilizationOrders* PerformanceMetric

Figure 24 provides an alternative view of the same data in Figure xx. To obtain this view, a bar chart and cross tabulation table were chosen. Once again, the Order Type and Sailor dimensions were dragged and dropped into the display. Further drill-down could be accomplished by double clicking on a bar or within the cross tabulation table. The arrows indicate scroll bars that can be used to see more of the data.

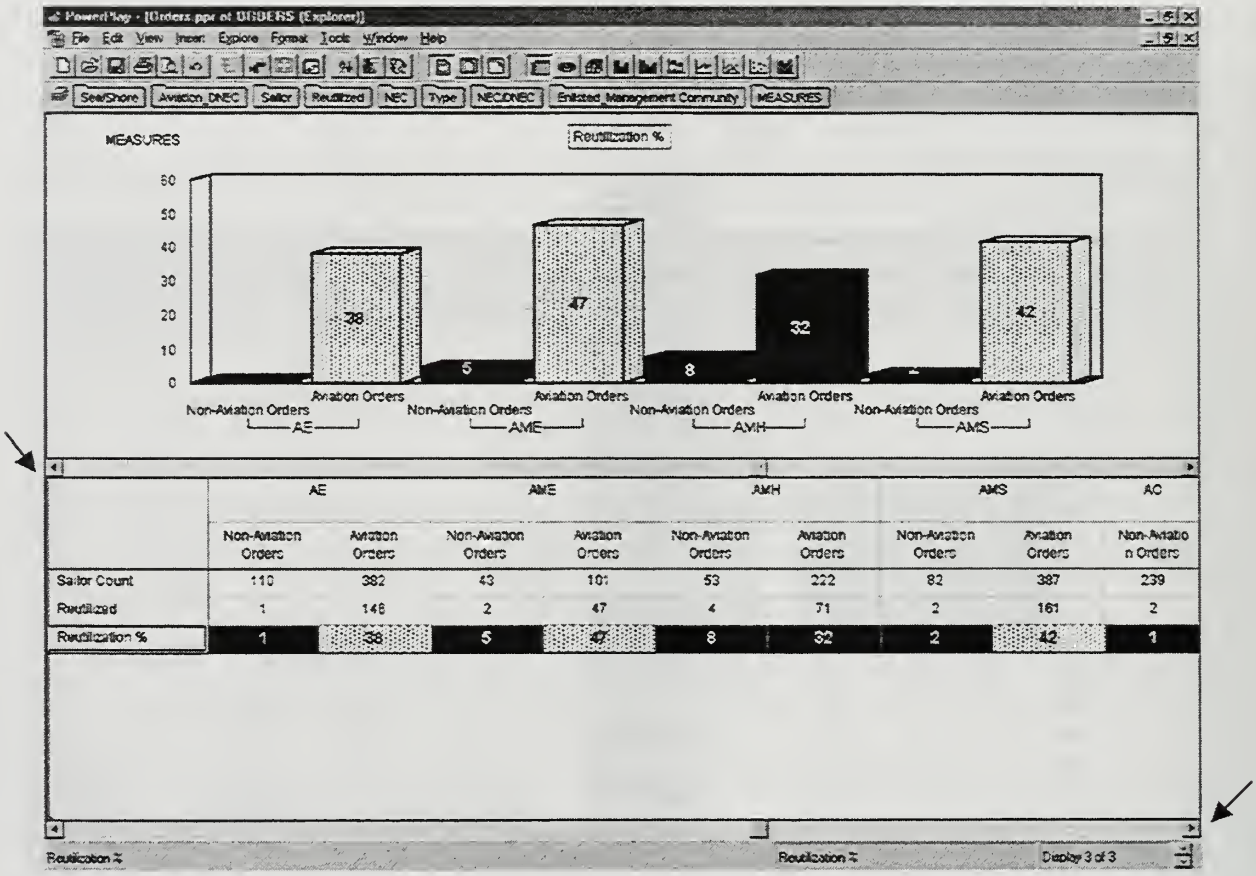


Figure 24. Alternate Rate View of *%NECReutilizationOrders* Performance Metric

Figure 25 provides NEC reutilization data from a command perspective. This provides a new view of the data by analyzing NEC reutilization rates by MCA. This view was obtained by dragging and dropping the Order Type and MCA dimensions into the display.

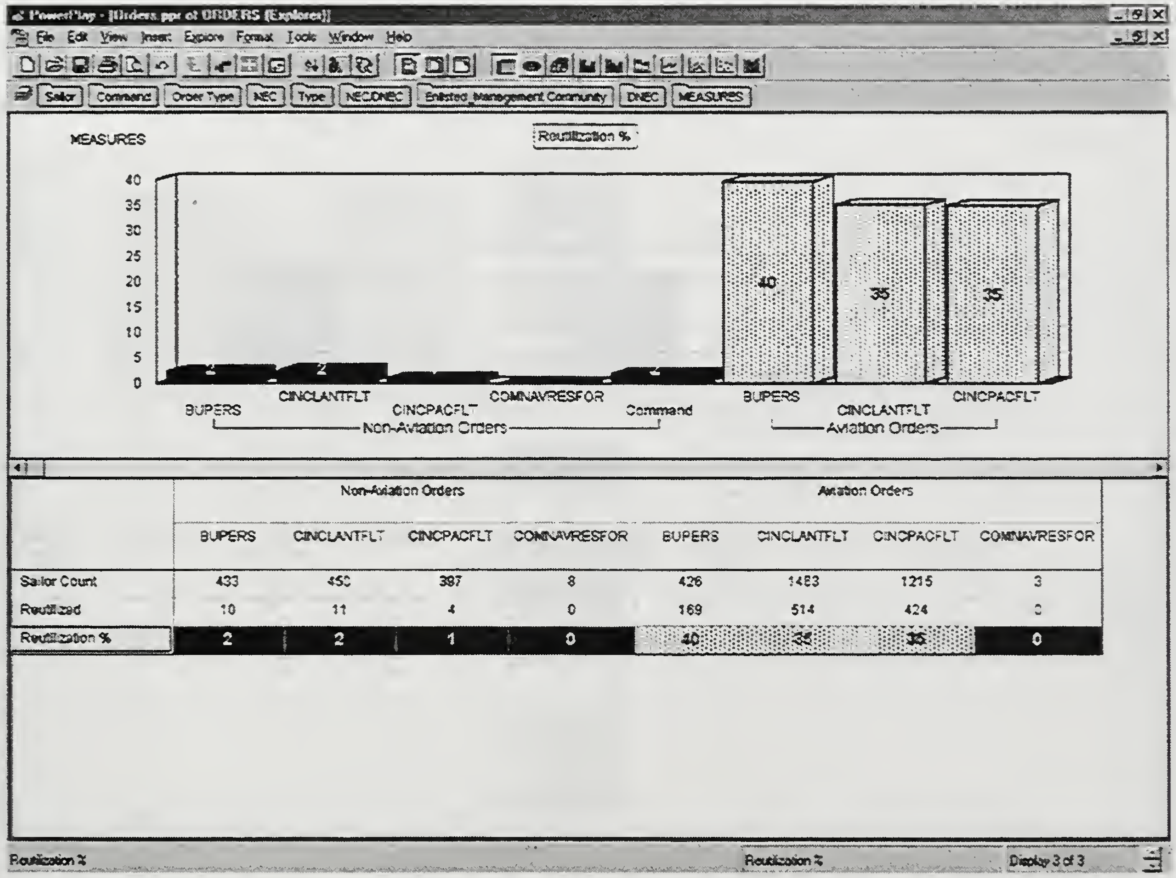


Figure 25. MCA View of *%NECReutilizationOrders* Performance Metric

Figure 26 is obtained by drilling down deeper into the data by double clicking the CINCPACFLT bar for personnel with Aviation Orders in Figure xx. This view demonstrates how continuously drilling deeper into the data can begin to reveal narrower views of the data.

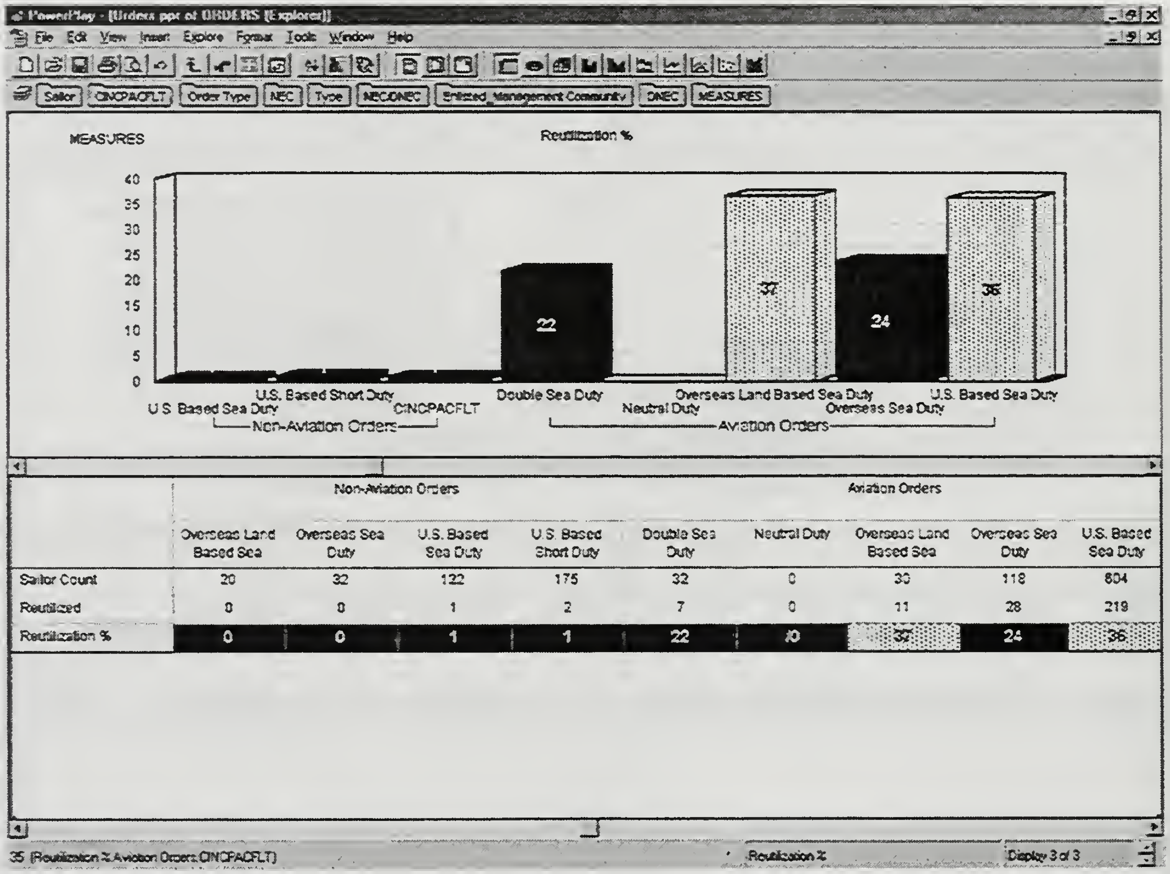
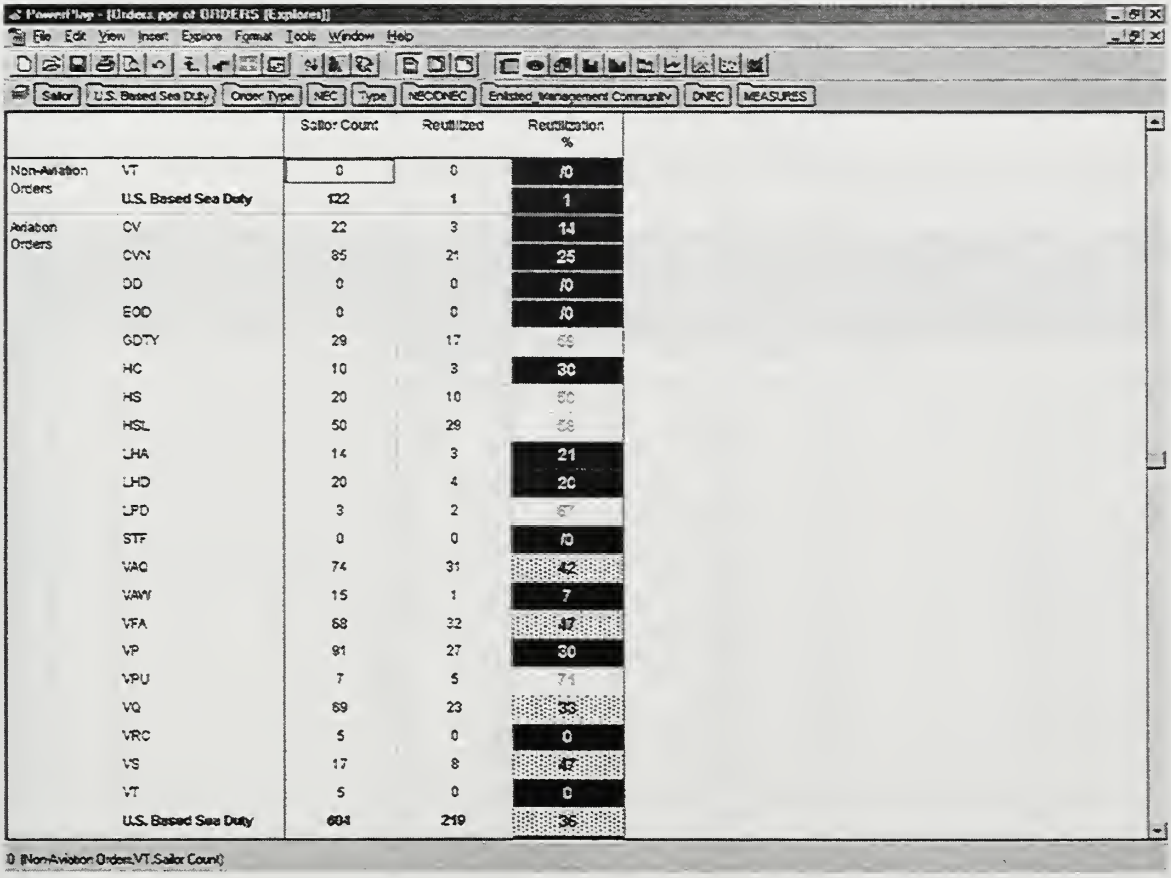


Figure 26. Sea/Shore View of *%NECReutilizationOrders* Performance Metric

Figure 27 is obtained by drilling down deeper into the data by double clicking the U.S. Based Sea Duty bar for personnel with Aviation Orders in Figure xx. Once again, this provides a narrower view of the data by drilling down. This view would help a personnel planner determine if NEC reutilization was equitable throughout the fleet.



		Sailor Count	Reutilized	Reutilization %
Non-Aviation Orders	VT	0	0	/0
	U.S. Based Sea Duty	122	1	1
Aviation Orders	CV	22	3	14
	CVN	85	21	25
	DD	0	0	/0
	EOD	0	0	/0
	GDTY	29	17	59
	HC	10	3	30
	HS	20	10	50
	HSL	50	29	58
	LHA	14	3	21
	LHD	20	4	20
	LPD	3	2	67
	STF	0	0	/0
	VAQ	74	31	42
	VAW	15	1	7
	VFA	68	32	47
	VP	91	27	30
	VPU	7	5	71
	VQ	69	23	33
	VRC	5	0	0
	VS	17	8	47
	VT	5	0	0
	U.S. Based Sea Duty	604	219	36

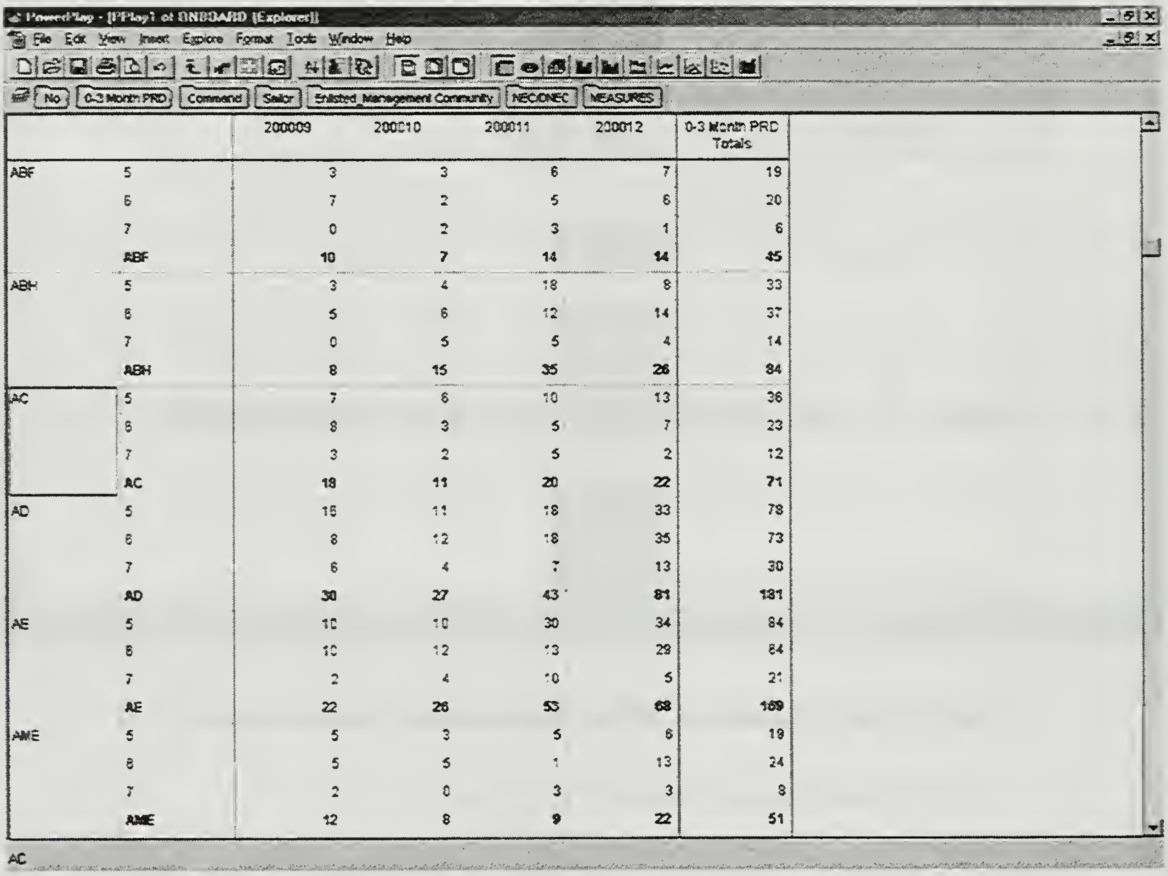
Figure 27. Platform View of *%NECReutilizationOrders* Performance Metric

2. %NECReutilizationOnboard Analysis

Figures 28 through 31 provide a sampling of PowerPlay™ screen shots of the *%NECReutilizationOnboard* performance metric. Data can be analyzed for personnel under orders, not under orders or a combination of both. In addition to tracking NEC reutilization rates for Rates, Paygrades, Commands, etc., this metric also gives planners the ability to look forward using PRD's for personnel not currently under orders. For

example, a planner can find which individuals are due to rotate in 0-3, 3-6, 6-9 or 9-12 month intervals.

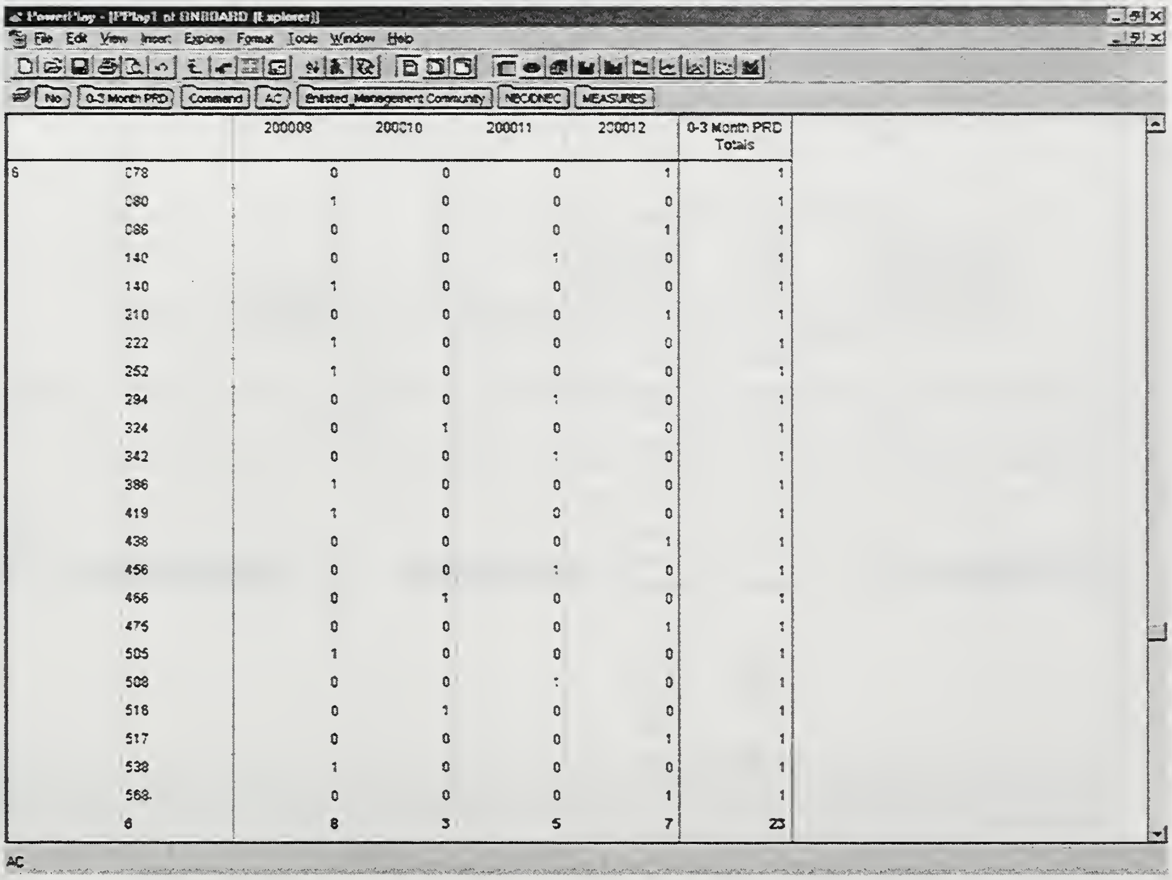
Figure 28 provides a cross tabulation table of the number of sailors due to rotate by the Rate and Paygrade dimensions. This projection covers the current month of September (200009) through December (200012). Drilling down deeper is easily accomplished by double clicking a Paygrade. Additional information could be analyzed by dragging and dropping those dimensions onto the display.



		200009	200010	200011	200012	0-3 Month PRD Totals
ABF	5	3	3	6	7	19
	6	7	2	5	6	20
	7	0	2	3	1	6
	ABF	**10**	**7**	**14**	**14**	**45**
ABH	5	3	4	18	8	33
	6	5	6	12	14	37
	7	0	5	5	4	14
	ABH	**8**	**15**	**35**	**26**	**84**
AC	5	7	6	10	13	36
	6	8	3	5	7	23
	7	3	2	5	2	12
	AC	**18**	**11**	**20**	**22**	**71**
AD	5	16	11	18	33	78
	6	8	12	18	35	73
	7	6	4	7	13	30
	AD	**30**	**27**	**43**	**81**	**131**
AE	5	10	10	30	34	84
	6	10	12	13	29	64
	7	2	4	10	5	21
	AE	**22**	**26**	**53**	**68**	**169**
AME	5	5	3	5	6	19
	6	5	5	1	13	24
	7	2	0	3	3	8
	AME	**12**	**8**	**9**	**22**	**51**

Figure 28. PRD/Rate View of *%NECReutilizatioOnboard* Performance Metric

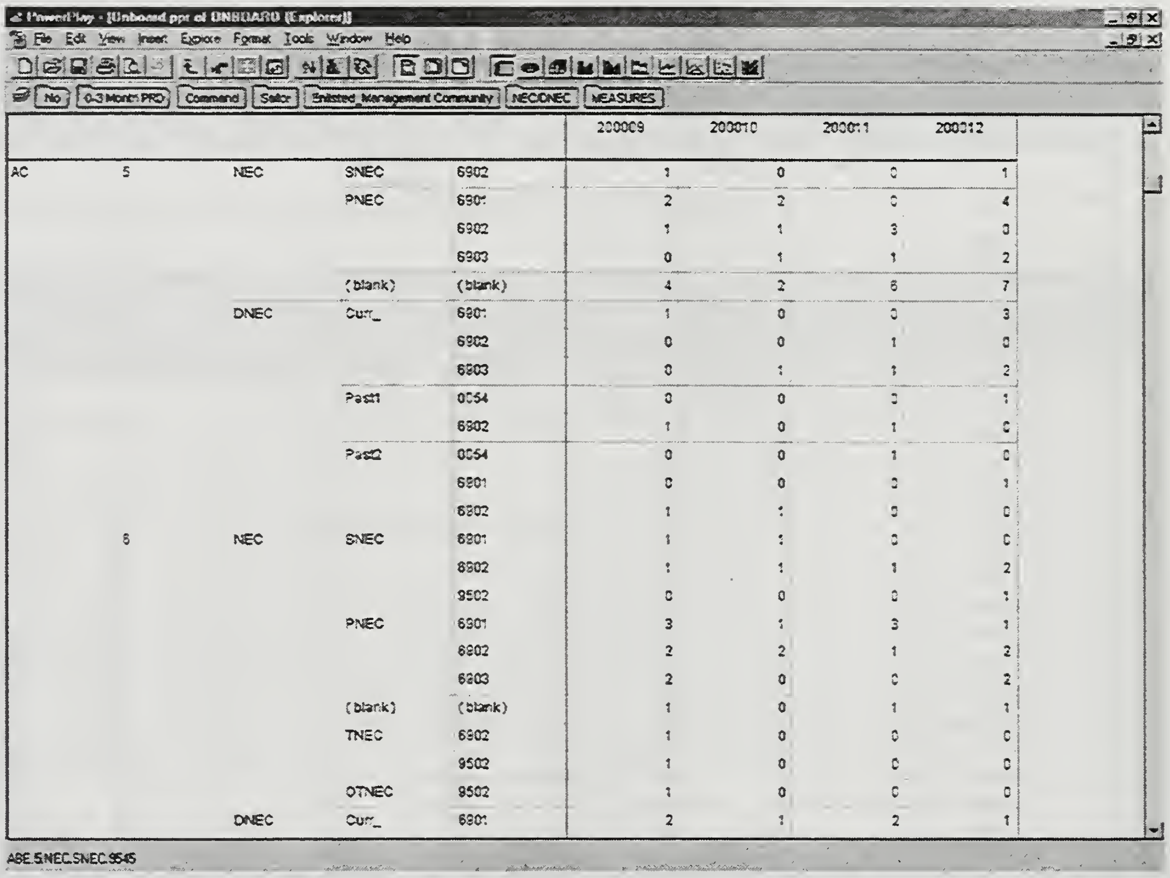
Figure 29 was obtained by double clicking the AC6 Rate and Paygrade in Figure xx. This adds the SSN information. The SSN's have been edited to 3 numbers. Like all views, this could be saved as an HTML document or be included in a more formalized report for easy distribution.



		200009	200010	200011	200012	0-3 Month PRD Totals
6	078	0	0	0	1	1
	080	1	0	0	0	1
	086	0	0	0	1	1
	140	0	0	1	0	1
	140	1	0	0	0	1
	210	0	0	0	1	1
	222	1	0	0	0	1
	252	1	0	0	0	1
	294	0	0	1	0	1
	324	0	1	0	0	1
	342	0	0	1	0	1
	386	1	0	0	0	1
	419	1	0	0	0	1
	438	0	0	0	1	1
	456	0	0	1	0	1
	466	0	1	0	0	1
	475	0	0	0	1	1
	505	1	0	0	0	1
	508	0	0	1	0	1
	516	0	1	0	0	1
	517	0	0	0	1	1
	538	1	0	0	0	1
	568	0	0	0	1	1
	6	8	3	5	7	23

Figure 29. PRD/SSN View of *%NECReutilizationOnboard* Performance Metric

Figure 30 was obtained by adding Rate, Paygrade, NEC, DNEC and 0-3 month PRD date dimensions. Adding SSN's at this point would provide a complete NEC history for each sailor for the PRD projection date desired.



					200009	200010	200011	200012
AC	5	NEC	SNEC	6902	1	0	0	1
			PNEC	6901	2	2	0	4
				6902	1	1	3	0
				6903	0	1	1	2
			(blank)	(blank)	4	2	6	7
		DNEC	Curr_	6901	1	0	0	3
				6902	0	0	1	0
				6903	0	1	1	2
			Past1	0054	0	0	0	1
				6902	1	0	1	0
			Past2	0054	0	0	1	0
				6901	0	0	0	1
				6902	1	1	0	0
	6	NEC	SNEC	6901	1	1	0	0
				6902	1	1	1	2
				9502	0	0	0	1
			PNEC	6901	3	1	3	1
				6902	2	2	1	2
				6903	2	0	0	2
			(blank)	(blank)	1	0	1	1
			TNEC	6902	1	0	0	0
				9502	1	0	0	0
			OTNEC	9502	1	0	0	0
		DNEC	Curr_	6901	2	1	2	1

Figure 30. PRD View of *%NECReutilizationOnboard* Performance Metric

Figure 31 presents another view of the data, this time using the ECM and MCA dimensions. This view provides NEC reutilization for all personnel, regardless of order status.

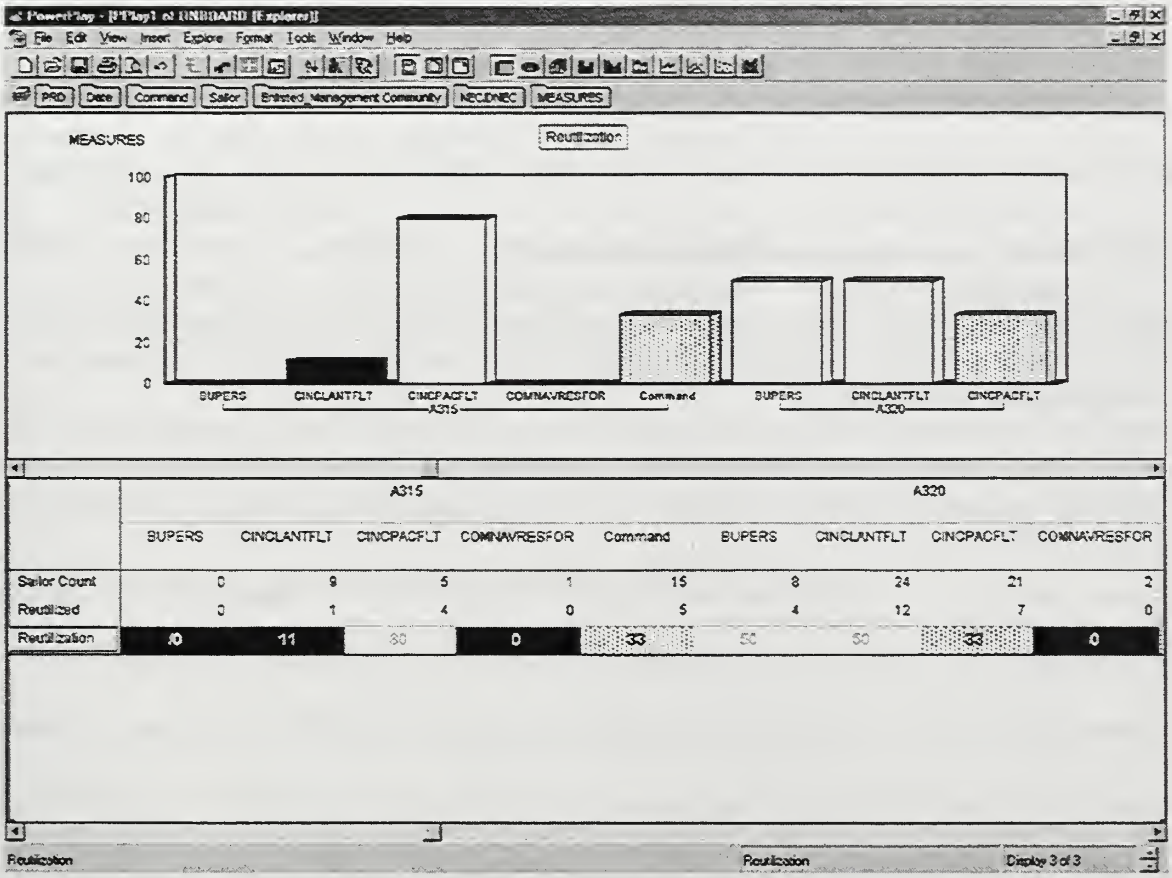


Figure 31. ECM/MCA View of *%NECReutilizationOnboard* Performance Metric

F. SUMMARY

Our goal in this chapter was to use an OLAP tool with a relational data mart to solve a specific real world problem concerning enlisted personnel assignment in the Navy. As is evident, an almost infinite number of data views are available for the two performance metrics we developed. We feel confident that the prototype can provide enlisted planners the knowledge they need to make more effective assignment decisions. We also believe this solution is applicable to a wide range of problems that require

unifying legacy-based systems into an integrated database-driven system. In addition, we believe use of the full spectrum of Cognos reporting products would provide a more polished reporting capability. We think the prototype shows great promise and can be used as starting point for further research and work in this area.

Chapter VI presents further arguments that the data warehouse approach is a better way to reengineer the legacy-based enlisted manpower systems than rebuilding those systems from scratch as database applications. In addition, we will discuss data quality issues as they pertain to migrating operational data from legacy-based systems to a relational data warehouse.

VI. DATA QUALITY

"Most information reengineering initiatives will fail due to lack of data quality." (Gartner Group, 2000)

"Gaining and sustaining information superiority requires DoD to field information systems that are interoperable at the data level." (Money, 2000)

A. INTRODUCTION

In November 1999 the DONCIO established the Data Management & Interoperability Integrated Product Team (DMI-IPT) to address data management and data quality issues at the enterprise-level. The DMI-IPT is a collaborative effort by DON organizations to address the current situation of independent data management strategies and propose an integrated enterprise approach. (DONCIO, 2000)

We believe that a relational data warehouse that incorporates OLAP tools is well suited to meet the DONCIO's integrated enterprise-level approach to data management. In addition, we believe it is a better way to re-engineer the enlisted manpower legacy systems than rebuilding those systems from scratch into database applications. This is because it allows incremental deployment, a myriad of data warehouse and OLAP applications are available, it is feasible, it does not require parallel operation of the systems during transition and an entirely new system would be too complex and costly.

However, for the warehouse approach to be implemented successfully, data quality initiatives must be given a high priority. Many people use the term "garbage in, garbage out" to describe the resultant negative effects of inaccurate data being entered into a database system. This phenomenon clearly applies to quality in terms of migrating operational data into a data warehouse or data mart for subsequent analysis with an

OLAP tool. This flows from the fact that a common thread among failed projects is that operational source data was not properly evaluated and prepared prior to loading into a data warehouse. In order to bring awareness to this issue, this section provides a high-level overview of data quality. We begin by providing a general discussion of data quality. We then present high-level views of the issues inherent to achieving data quality; source system analysis, data migration, metadata, Web technology, data architecture and data management approaches.

> The single most important success factor for data warehousing is the **quality of information** provided to data warehouse users...A data warehouse that contains trusted, strategic information becomes a valuable enterprise resource for decision makers at all organizational levels...If it's users discover that it contains bad data, the data warehouse will be ignored and fail.... (Perkins, 1998)

B. WHAT IS DATA QUALITY?

One perspective of data quality emanates from a user perspective rather than technical considerations such as interoperability or integration. After all, it is the users that ultimately will decide if data is of a high quality or not, hence, we adopt the following definition. Data quality is "consistently meeting knowledge worker and end-customer expectations." (English, 2000) Table 17 presents a list of data quality characteristics to consider when evaluating whether the information will meet end-user requirements. This is a useful perspective for IT personnel to keep in mind continuously since it can easily be lost when designing and building complex systems.

Characteristic	Knowledge Worker benefits
The right data	The data I need
Completeness	All the data I need
The right context	I know the meaning
The right accuracy	I can trust the data
No uncontrolled redundancy	I have a single version of the truth
The right format	I can use it easily
The right time	When I need it
The right place	Where I need it
The right purpose	I can accomplish our objectives

Table 17. Data Quality Characteristics (English, 2000)

C. SOURCE SYSTEM ANALYSIS

The DON DMI-IPT established the Authoritative Data Sources sub-team to work on data source issues.

> The focus of this group is the myriad of issues associated with defining and identifying the primary sources of data within communities of interest. The inability to identify authoritative data sources was found to be a major factor in the ability to respond quickly to year 2000 data remediation efforts. The subject of authoritative data sources is yet another key element to a strategy that supports horizontal integration of information across different functional areas. (DONCIO, 2000)

Source system analysis is required because source data is often maintained in disparate databases that contain fields that have taken on different meanings over many years. Making sense out of the fields is just one of many problems that can plague the early phase of any data warehouse project. To avoid this situation requires that the true meaning of all the data be meticulously evaluated prior to migration. (Watterson, 1998)

The assessment of source systems involves determining *what* source systems to pull data from and *how* this should be done. The objective of source system assessment is to determine the best source of data (often called the *system of record*) for each fact

and dimension in a data warehouse as well as the best method for extracting this data on a regular basis for populating the data warehouse. In addition, it must be decided how often to extract source system data to a data staging environment for data preparation and loading into the data warehouse. This can be difficult because it must be determined how this can be done without causing a severe performance impact on the source system. A data quality assessment is often done in conjunction with a source systems assessment to determine the general level and reliability of the source data. Data sampling is often used to determine the overall level of quality that exists within proposed systems.

D. EXTRACTION, TRANSFORMATION & LOADING (ETL)

The goal of ETL is to move operational data into a data warehouse or data mart while ensuring data quality. Activities conducted during ETL include *data preparation*, *data cleansing*, *data transformation* and *data integration*. During our literature search, we have found these terms are often used interchangeably. The following definitions provide the distinguishing characteristics of the terms:

- **Data preparation** - A general concept that can involve data cleansing, data migration, data transformation and/or data integration.

- **Data cleansing** - Usually associated with the term "dirty data." Dirty data includes the following problems associated with data: dummy values being entered, absence of data, multipurpose fields, cryptic data, contradicting data, inappropriate use of fields, violation of business rules and/or reused primary keys. (Moss, 1998)

- **Data transformation** - Tools that extract data from operational sources, clean it, and load it into a data warehouse.

- **Data integration** – "Data integration is an interoperable and extensible framework to easily share one common integrated view of the information as it flows throughout the enterprise value chain." (Merant, 1998) Data integration involves the process, issues and problems associated with integrating data from multiple operational databases into a data warehouse or

data mart. This often involves implementing data quality through metadata (See discussion of metadata in Section E below).

Typical problems associated with data integration include data that should be related, but cannot be and data that are inadvertently related but should not be. [Moss, 1998] The former issue highlights a problem of the current enlisted database system. Ideally, a personnel management system should integrate personnel and billet data so manpower planners get a complete picture of available personnel and manpower requirements. In other words, a billet should be related to a sailor. However, in the Navy enlisted system, personnel data is maintained in the EMF and billet data is maintained in the TFMMS. This system does provide a seamless link between the two.

The fact that data cleansing and metadata quality initiatives are necessary highlights another reason why a data warehouse reengineering approach is beneficial. Legacy-based database systems that have evolved over time with little or poor documentation are almost destined to have quality problems. If done correctly, ETL processes can correct the data quality problems associated with legacy-based database systems better than totally rebuilding new database applications. This is because the data warehouse approach allows a systematic deployment that can include risk management principles. For example, ETL, relational data warehouse and OLAP deployment can be implemented by proof-of-concept or pilot project first by incrementally deploying the system rather than whole-scale introduction of a new system. This approach enables unforeseen problems to be detected and solved more quickly and successfully then might otherwise be possible.

E. METADATA

The DON DMI-IPT established the Metadata Repository sub-team to develop the groundwork for providing awareness and access to data assets enterprise-wide.

> Accessibility to metadata is considered a key to information interoperability across business and warfighting systems and functions. (DONCIO, 2000)

> A *metadata repository* refers to the physical tables that will contain the metadata. The metadata repository supports every phase of development of the data warehouse, from requirements gathering, data model design, data mapping, user access, data warehouse maintenance, future warehouse development, and historical data needs. (Marco, 1998)

Because data is coming from many sources, each with their own data definitions, an enterprise-level metadata model must be created to provide a consistent view of the data. Without metadata, data quality in a data warehouse is very difficult to create or maintain. Metadata helps to catalog or define a particular resource, much like a library catalog card, see Figure 32. Metadata are used to provide documentation for data products. The primary purposes of metadata are:

- Facilitate the discovery of information.
- Assist in the management of information.

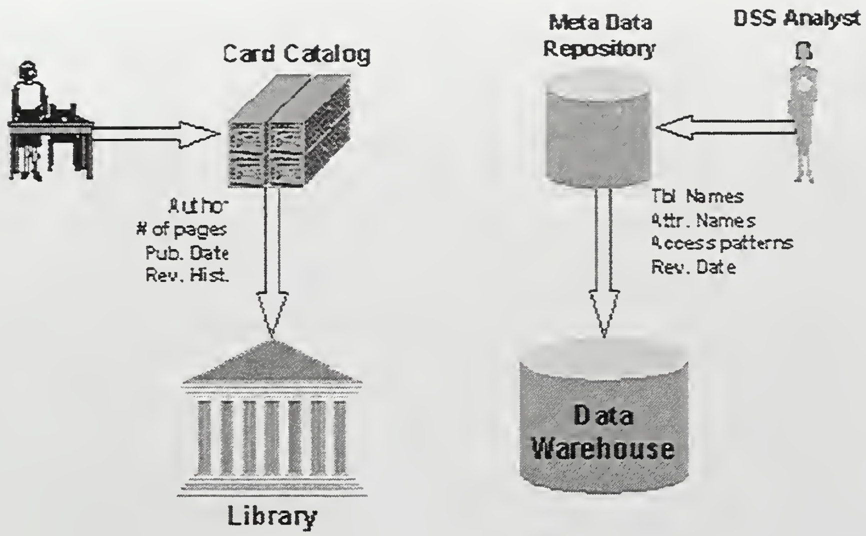


Figure 32. Metadata Repository (Marco, 1998)

We believe that DONCIO's clear emphasis on the importance of Metadata is highly supportable under the data warehouse/OLAP database-driven system we envision to replace the current enlisted personnel manpower system.

1. Microsoft Metadata Products

Microsoft has created the Microsoft Repository for defining an industry wide standard for the metadata model. The goal of Repository is to enable data warehousing products from different vendors to share information. Sharing and reuse of metadata requires an agreement on the metadata's structure and semantics. Microsoft calls the exact specification of such an agreement an *information model.* An information model ensures that all involved parties can encode and interpret the information that is exchanged. If a consensus can be developed among the vendors of these products, the new repository

extensions should help expand the data warehousing market by providing an open. common infrastructure across all data warehouse vendors.

Figure 33 shows a typical data warehouse architecture and how the Repository is used to store information about the entire warehouse

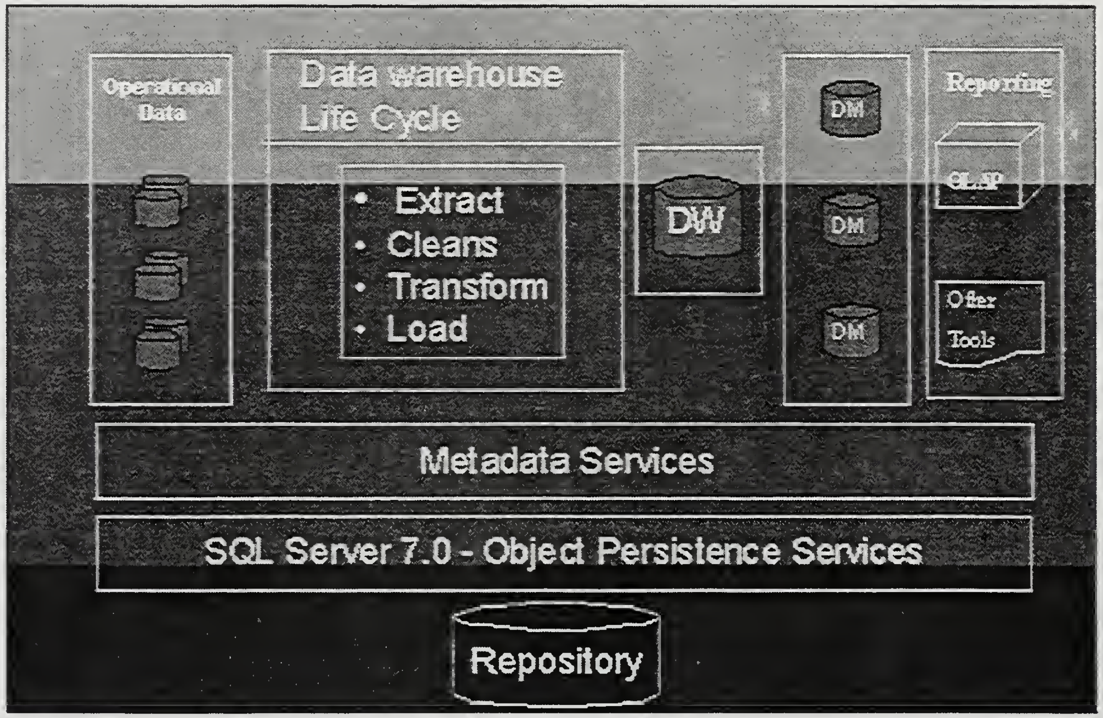


Figure 33. Data Warehouse Architecture & Repository (Cross, Rahimi, 1999)

Figure 34 shows how data warehouse components are integrated with the Repository, enabling extraction and storage of their metadata in the repository.

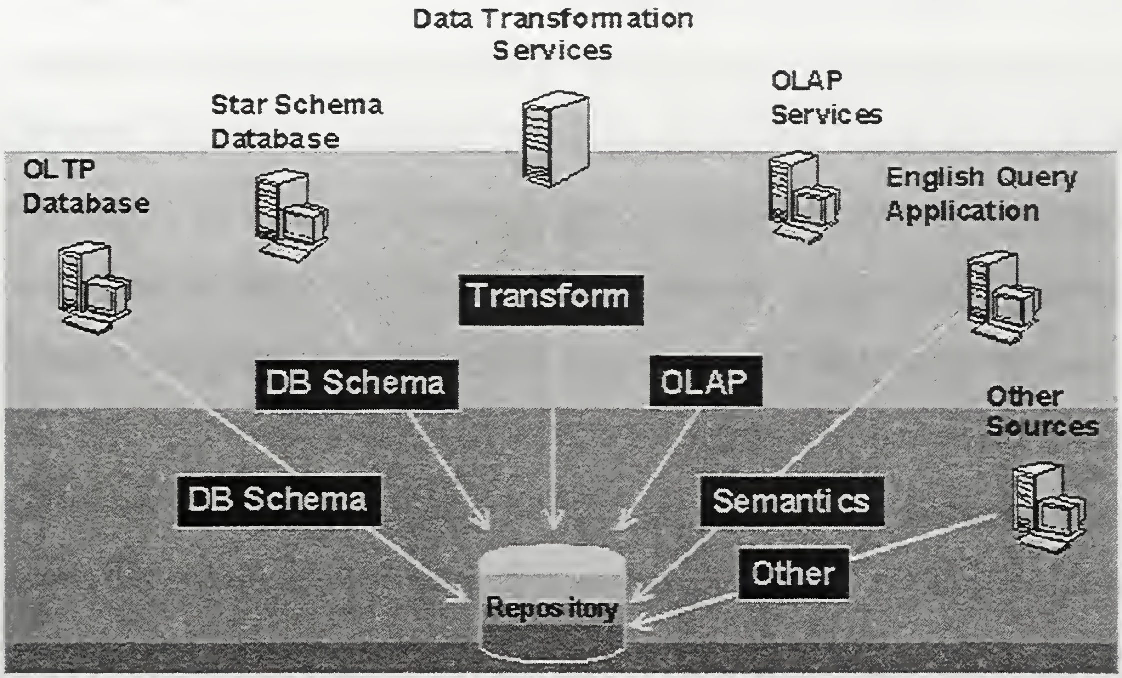


Figure 34. Repository Integration (Cross, Rahimi, 1999)

F. WEB TECHNOLOGY

Although not often thought of in terms of data quality, we believe the use of Web technology to deliver information to knowledge workers is a data quality issue. As discussed in Section B of this chapter, data quality includes providing easily accessible data to the right person, when and where it is needed. Without question, the use of Web technology to provide users access to information will be a critical component of any database-driven system. Data Warehouse and OLAP technology is now easily assimilated to a Web presence. Many OLAP applications are viewable within a Web browser and typically list predefined queries or reports that can be run against a multidimensional database.

1. Web Technology & Metadata

The combination of Web and metadata knowledge is a very important pairing of technologies. A current form of metadata with tremendous potential is Extensible Markup Language (XML). XML was developed by the World Wide Web Consortium (W3C) and works by adding "data tags" in HTML that describe the data (metadata) on a web page. This tag-based approach offers a flexible method to handle the metadata associated with information content (i.e., each tag name describes the characteristic of the data it specifies). XML allows the data tag to describe a data's meaning, not the location of the data. Therefore, data can be put in any order on a web page. Data quality would be unquestionably improved through the use of XML since the data tags can take disparate data and make it consistent.

G. ARCHITECTURE

"Architecture is a set of standards, guidelines, and statements of direction that constrain the design of solutions for the purpose of eventual integration." (Kesnor, 2000) In order to unify enlisted manpower legacy systems into an integrated database-driven system, the appropriate data architecture must be selected. To help address this issue, DMI-IPT established the Architecture & Standards sub-team to define the processes associated with the development of data architectures.

> Data architecture, as defined within DoD, is a framework for organizing the interrelationships of data, providing the incremental, ordered design and development of systems based on successively more detailed levels of data modeling. The process for developing data architectures includes the registration of data within existing systems and the development of standards that support the goals of data management. (DONCIO, 2000)

When choosing an enterprise-level database deployment strategy, it is critical that IT managers think in terms of *architecture.* Often times, database procurement decisions

are made under the guise of an IT architectural solution that would provide both immediate and long-term solutions. However, in reality many of these solutions are proprietary systems that provide only short-term fixes for a particular problem. Proprietary systems are typically the quickest to succumb to legacy-based status and often cause lock-in to a single source support vendor. One developer sums it up well by stating, "Evolving technology requires that organizations optimize for the life - not the birth - of a system." (Forrester Research, 2000)

In order to be truly effective, it is critical that the architecture not standardize on a product set (for example NT or Oracle) but rather on open standards such as ODBC, HTTP, CORBA, HTML, etc. Strategically leveraging open standard technologies will empower an organization to design and implement adaptable and unique solutions through the application of distributed components and services. The Navy's roadmap should follow a path founded upon extensibility and a fully scalable architecture that will provide the customization, functionality and platform independence required to distribute needs across systems and applications.

H. MANAGEMENT APPROACH

The fourth team established by the DMI-IPT is the Management sub-team.

> The Management sub-team was established to address the requirements associated with the production and use of data within functional activities and information systems. This sub team is also addressing commands' and individuals' roles and responsibilities and reviewing data management requirements that need to be included in DON and DoD budgeting and acquisition processes. (DONCIO)

The DMI-IPT Management sub-team's efforts focus on the importance of having a strategic management plan in place to achieve data quality. Data warehousing can affect the assessment and planning for future business activities as well as day-to-day

operations. It is important to understand the impact of the technology on the users and how it is being applied to try to overcome and avoid personnel problems. To ensure that the infrastructure and end-user issues are addressed as part of the overall delivery of the data warehouse, a change management assessment must be undertaken.

Ensuring data quality can involve very complex and time-consuming tasks that require the assistance of many skilled individuals with wide-ranging fields of knowledge. In fact, it is not uncommon for data quality initiatives to require 50 to 70 percent of a project's budget and labor in order to help guarantee success. A highly skilled team of experts must be in place to successfully implement a data warehouse while ensuring data quality. This team will likely include both in-house development staff and outside consultants. The following list offers a snapshot on the types of members who should make up the team: (Freeman, 1997)

- **Project manager** - Facilitate and supervise the activity.
- **Business sponsor** - Person held accountable for the projects success and who understands the strategic advantage provided by the data warehouse.
- **Business operations manager** - Keeps the project on track by knocking down roadblocks.
- **IT manager** - Has the overall vision and strategic plan, will enforce standardization and manage the technical challenges.
- **Database administrator** - Tunes the performance of the database.
- **Network administrator** - Optimizes bandwidth usage and ensures the data warehouse is available to users.
- **Data administrator** - Loads and cleans data.
- **Hardware administrator** - Optimizes the performance of all tools on the hardware platform.

- **Decision support administrator** - Bridges the gap between users and the IT department.

- **Users** - Should include casual users to power users to ensure that customers will be satisfied with the final product.

THIS PAGE INTENTIONALLY LEFT BLANK

VII. CONCLUSIONS

A. SUMMARY

A well-implemented, integrated database-driven information system can have a major impact on an organization's ability to quickly and effectively make quality decisions based on accurate and reliable data. Data warehousing and multidimensional analysis of data plays a key role in such a system. Many organizations are now beginning to realize the benefits of this technology, and this sector of the IT industry is currently enjoying high growth.

We have presented a strong case that unifying Navy enlisted manpower legacy-based systems into an integrated database-driven system that includes a relational data warehouse and OLAP tool can offer dramatic improvements in business intelligence. We also believe that this approach is a better way to re-engineer the enlisted manpower legacy systems than rebuilding those systems from scratch into database applications. The data warehouse approach has the following advantages:

- A data warehouse and OLAP solution can be deployed incrementally without interrupting the operation of existing legacy systems. Further, the many homegrown systems such as ODIS and EAIS can eventually be discarded in favor of integrated reporting tools such as those provided by OLAP products.

- This approach is eminently feasible as demonstrated by our data warehouse and OLAP prototypes. The underlying database structures are not very complex. Our prototype has shown that the EMF file is the basic source database. Other files representing the billet side of the Navy can be added quite easily as we have shown.

- Almost all contemporary data warehouse and OLAP tools are fully functional in an Internet or intranet environment. This is an important feature since providing easy access to dispersed users is a significant facet of worker empowerment.

- Designing and implementing new database applications to replace the legacy systems would probably be extremely costly and complex. Certainly, the data

warehouse can be developed for considerably less money that a complete relational "makeover" of existing systems. In addition, many of the ancillary query systems would no longer be needed, resulting in additional significant savings.

- The data warehouse approach would not require a period where both systems would have to be maintained in parallel as a transition from the current to the new system took place. This is one of the major obstacles to rebuilding the legacy systems "from scratch".

- A data warehouse approach is imminently scalable to meet growing demand over time.

Data warehousing and data quality initiatives are wholly consistent with the DONCIO's emphasis toward developing integrated and interoperable database platforms across functional areas. This may signify the most important advantage. Designing and implementing a relational data warehouse with data quality in mind will ensure that the DMI-IPT core initiatives of source system analysis, metadata, architecture and management are addressed.

We believe that the combination of relatively low cost development and decision-making advantages of the data warehouse and OLAP approach versus the operational capabilities of the current enlisted manpower system speaks for itself. There is little doubt in our minds that this approach will provide enlisted planners a solid foundation to make knowledge-based decisions. Clearly, this proven technology provides a migration path that will benefit the Navy for years to come.

B. PROTOTYPE DELIVERABLE

On 07 September 2000 the relational data mart and OLAP prototype was demonstrated for the Branch Head, Aviation Enlisted Assignments (PERS-404). As a result of this meeting, minor changes were made to the prototype to optimize the system for NEC reutilization and PRD analysis. The Microsoft Access data mart, Cognos

PowerPlay OLAP application and OLAP models were then loaded onto a laptop computer to provide PERS-404 a fully functional analysis tool.

C. WEB-TARGET

Web-Target is a Web-based OLAP application designed to provide similar functionality to what we have demonstrated in our approach. The Navy Personnel Research Studies and Technology (NPRST) group sponsors Web-Target, which is billed as a **WebIntelligence** system that provides the functionality of professional decision support tools over the Web. To accomplish this, it provides many pre-defined data dimensions that can be used for analysis (Figure 35). WebTarget supports the basic concepts we espouse in this thesis although our approach captures more details about aviation NEC reutilization in particular. Since WebTarget is already web-enabled, it has wider accessibility than our system although the data is not refreshed as frequently. It would appear that WebTarget could serve as a first step in the direction we recommend.

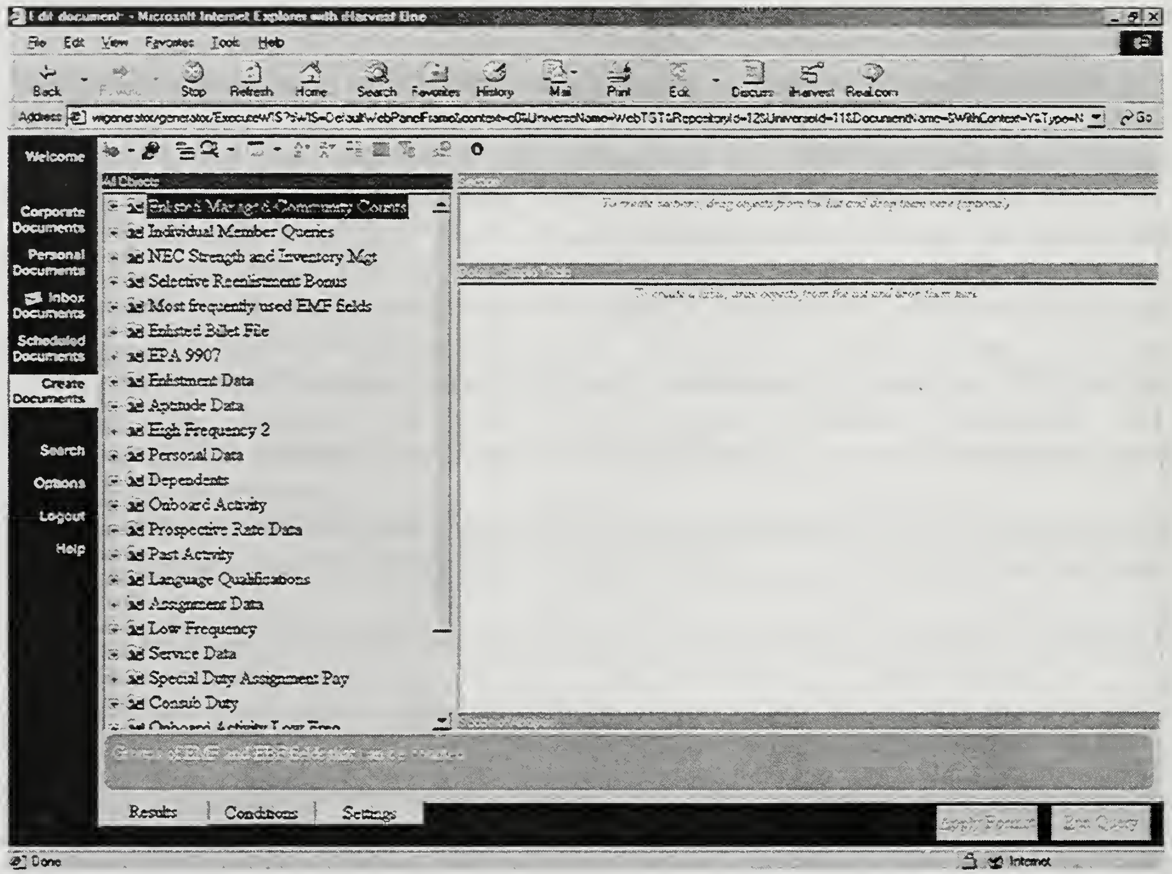


Figure 35. Web-Target

LIST OF REFERENCES

[Citation by name: in alphabetical order:]

Adelman, S. and Oates, J., Data Warehouse Project Management, 2000.

Angoss Software Corporation, A New Generation of Data Mining Technologies, 1998.

Clemins, A. Admiral, IT 21: The Path to Information Superiority, 1997.

Cross, P., & Rahimi, S, Leveraging the Microsoft Repository in Your Warehouse, 1999.

Dolk, D. R., Integrated Model Management in the Data Warehouse Era, 1999.

DONCIO, www.doncio.navy.mil, 2000.

English, L. P., Information Quality Management: The Next Frontier, 2000.

Forrester Research, 2000.

Freeman, E., Birth of a Terabyte Data Warehouse, 1997.

Gartner Group, www.gartnergroup.com, 2000.

Gaskin, B., Realizing the Strategic Value of Data Warehouses, 1999.

Haiston, M, The Real-Time Enterprise: The Real-Time Data Warehouse: The Next Stage in Data Warehouse Evolution, 1999.

Hurwitz Group, Scaleable Enterprise OLAP Requirements: Data depth and Breadth, 1998.

Information Discovery Inc., A Characterization of Data Mining Technologies and Processing, 1998.

Inhoff, C., Ensuring Data Quality Through Data Stewardship, 1998.

Inmon, B., Data Mart Does Not Equal Data Warehouse, 1999.

Kachur, R., Source System Assessment for a Data Warehouse, 2000.

Kesnor, R., IT planning and procurement: The underlying architectural process, 2000.

Marco, D., Implementing Data Quality Through Metadata, 2000.

Marco, D., Managing Metadata Today and into the 20th Century, 1998.

Merant Solutions Inc., Data Integration: Improving the Effectiveness of Data Across the Enterprise Value Chain, 1998.

Microsoft Corporation, SQL Server 7.0 Features Guide, 2000.

Microstrategy Inc., The Case for Relational OLAP, 1997.

Money, A., 7 February 2000 memorandum, ASD (C3I), 2000.

Moss, L., Data Cleansing: A Dichotomy of Data Warehousing, 1998.

OLAP Council, OLAP Council White Paper, 1997.

OLAP Report, Market Segment Analysis, 2000.

OLAP Report, What is OLAP, 2000.

Parsaye, K. & Petrie, D., Measuring the Value of Mined Information, 1998.

Parsaye, K., Data Mines for Data Warehouses, 1998.

Perkins, A., Critical Success Factors for Data Warehouse Engineering, 1998.

Pilot Software, An Introduction to Data Mining: Discovering Hidden Value in your Data Warehouse, 1997.

Ramakrishnan, R. and Gehrke, J., Database Management Systems, 2000.

Senge, P. M., The Fifth Discipline, The Art & Practice of The Learning Organization, 1990.

SPSS, Data Mining Techniques, 2000

Sun Microsystems, Inc., Data Warehousing – Scaling the Data Refinement, 1998.

Watterson, K., The New Face of Data Warehousing, 1998.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center 2
   8725 John J. Kingman Road, Ste 0944
   Fort Belvoir, VA 22060-6218

2. Dudley Knox Library 2
   411 Dyer Road
   Monterey, CA 93943-5101

3. Dean Dan Boger 1
   Code IW
   Naval Postgraduate School
   Monterey, CA 93943-5118

4. Professor Dan Dolk, Code IS/DK 1
   Naval Postgraduate School
   Root Hall, Rm. 229A
   Monterey, CA 93943-5118

5. Lecturer Julie Filizetti, Code IS/DK 1
   Naval Postgraduate School
   Monterey, CA 93943-5118

6. CDR Tim Ferree 1
   Navy Personnel Command (NPC)
   Aviation Enlisted Assignments (PERS-404)
   5720 Integrity Drive
   Bldg. 791, Rm. H202A
   NSA Mid South
   Millington, TN 38055-4040

7. LCDR Douglas J. Conde.. 2
   143 General Stonewall Jackson Blvd.
   Harpers Ferry, WV 25425

8. LCDR Cassandra A. Crownover 2
   2248 Grand Avenue
   Deland, FL 32720